

Cairo University

Institute of Statistical Studies and Research

**The 53rd Annual Conference on Statistics, Computer
Sciences and Operations Research**

COMPUTER SCIENCE

3-5, Dec ,2018

Index
Computer Sciences

1	Ontology Learning Process from Arabic Text: A Comprehensive Survey Mariam Muhammad, Nesrin Ali, and Mervat Gheith	1-23
2	Vortex Swarm Optimization Ahmed Sabry A. Rahman and Hesham A. Hefny	24-41
3	Design Obfuscation for IP Security Applications Ahmad M.H.Albatsha	42-47
4	A Comprehensive E-government Adoption Literature Review Abeer Mosaad Ghareeb, Nagy Ramadan Darwish, and Hesham A. Hefny	48-64
5	Optimization of Water Distribution Networks using PSO Algorithm Amira F.M., A.A. Abd El-Aziz, Hesham A.H and Ahmed. F.A	65-83
6	Deep Learning Based Diabetic Retinopathy Diagnosis: A Survey Ahmed M.Abdo, Ahmed H. Asad and Hesham A. Hefny	84-100
7	Ontology based Intelligent Tutoring Model for Analysis and Generation of Arabic Verbs and Derived Nouns Laila Ibrahim, Mostafa Ezzat, Mervat Ghith, and Magdy Aboul Ela	101-115
8	A Survey of Sign Language and Machine Translation Systems Mohamed Amin, Ammar Mohamed and Hesham Hefny	116-143
9	An Efficient Ambulance Routing System for Emergency Cases Based on Dijkstra's Algorithm, AHP, and GIS Sayed Ahmed, Romani Farid Ibrahim, and Hesham A. Hefny	144-166
10	Survey on Social Network Analysis M.M. Abo Khedra, A.A. Abd El-Aziz and Hesham A. Hefny	167-178
11	A Deep Learning Approach for Text Generation Belal Saeed Mahmoud, Mohamed Ezzeldeen Saleh and Ammar Mohammed	179-189
12	AFSCloud Uniform Restful API Data Access for Cloud Data Store Sherif Younis Saleh, Fatma A. El-Licy and Akram Salah	190-200
13	A Survey of Quantum Data Compression, Methods, and Approaches Mohamed Nehad, Fatma A. El-Licy and Akram Salah	201-213
14	The Leap from Digital Quantum Simulators, to Quantum Supremacy Hisham Nabil, Fatma A. El-Licy and Akram Salah	214-226
15	Improving Storage Capacity of QR Code: A Survey Islam M. El-Sheref, Ahmed H. Asad and Fatma A. El-Licy	227-248
16	Record Linkage Techniques for Handling the Volume Property of Big Data Randa M. Abd El-Ghafar, Ali H.El-Bastawissy, Eman S.Nasr and Mervat H. Gheith	249-261
17	Overview of Integrating Granular Computing with Big Data Osama Sayed Abdel-Rahman, and Hesham Ahmed Hefny	262-273

Ontology Learning Process from Arabic Text: A Comprehensive Survey

Mariam Muhammad¹

Nesrine Ali²

Mervat Gheith³

Abstract. Ontology learning is an important process for the knowledge representation in many fields especially in the Semantic Web. The success of the Semantic Web depends on the quality of its underlying ontologies. This paper presents a comprehensive survey of research on the Ontology Learning for Arabic texts. Most of the research works focused on three main issues: extracting the terms, extracting the semantic relations, or building the ontology from the Arabic text. In this paper, first we underline the Arabic challenges that were the reasons to develop few Arabic Ontology Learning systems. Second we make a research comparison based on the techniques used and their results. Finally, we pointed comments and limitations of research works on Arabic Ontology Learning.

Keywords: Arabic Ontology Learning, Term Extraction, Semantic Relation Extraction.

1 Introduction

The World Wide Web is a huge repository of data. The majority of this data is unstructured. The huge data is difficult to process by humans, while the unstructured format is difficult to understand by machines. Tim-Berners Lee invented the Semantic Web to handle the challenges of the huge data and the unstructured format [Berners-Lee et al., 2001]. He said that "The Semantic Web is an extension of the current web in which information is given well-defined

¹ Department of Computer Science, Institute of Statistical Studies and Research, Cairo University, Egypt.
eng.maryamadel@yahoo.com

² Department of Information Systems and Technology, Institute of Statistical Studies and Research, Cairo University, Egypt. nesrinealiazim79@hotmail.com

³ Department of Computer Science, Institute of Statistical Studies and Research, Cairo University, Egypt.
mervat_gheith@yahoo.com

meaning, better enabling computers and people to work in cooperation". Ontology can be considered as a gateway to achieve the aim of the semantic web where the data can be represented in a way that enables machines to understand its meaning. Also, the ontology allows data to be shared and reused [Somodevilla et al., 2018].

Gruber defined the ontology as “a description of the concepts and relationships” [Gruber, 1993]. In addition to the ontology importance in the semantic web, it is also very important in more of fields such as Data Integration and Interoperability, Machine Translation, and Information Retrieval. In Data Integration, the ontology can be used as a semantic reference to several information systems. Using the ontology with the Machine Translation task is by finding the exact mapping of concepts across languages. But in the Information Retrieval task, it is used to enrich queries and improve the quality of the results, i.e. meaningful search rather than string-matching search. Due to its importance for the semantic web and other fields, the process of building ontologies is important. The success of the Semantic Web depends on the quality of its underlying ontologies. Ontologies can be built either manually, semi-automatically or automatically. Building the ontology manually needs to lots of efforts, time-consuming, and it has error-prone [Hazman et al. 2011]. To overcome these limitations, several researchers tried to build the ontologies by using semi-automatic or automatic methods. The process of building the Ontology semi-automatically or automatically is called “Ontology Learning” [Barforush, 2012]. Extracting semantic relations from the text is one of the most important phases in the Ontology learning process. It is essential in all Ontology learning phases starting from building Ontology from scratch down to populating and enriching existing ones.

In this paper we overview the research that works on Ontology Learning for Arabic text and their results, we found that the research focused on Extracting the terms, Extracting the semantic relations and building the Ontology automatically.

The outline of this paper is as follows: a background about the Ontology and the Ontology Learning are presented in section 2. The researches on the Term Extraction are reviewed in section 3 and the semantic relation extraction is reviewed in section 4. In section 5 we overviewed approaches and techniques for building the ontology automatically for Arabic text. We explained our comments and limitation of previous work in section 6. Finally the conclusion of this paper is in section 7.

2 Background

In this section, the Ontology and the Ontology Learning will be discussed.

2.1 Ontology

Ontology is a structure that describes set of terms used in the specific domain and the relationships that may exist between these terms. Ontology consists of individuals, classes and subclasses, attributes, and relations. The individuals in ontology are instances of the concepts or objects such as people, animals and plants as well as abstract individuals such as numbers and words. Ontology may have no individuals. Classes are sets of objects describe by set of attributes. Classes may classify individuals with help of these attributes. Some examples of classes are Person, car, Thing, etc. Attributes are properties or characteristics that classes can have. For example, (name, age and height) are properties of a person class or object. Relationships are links between objects in ontology that detect how objects are related to other objects such as (part-of, has-a, is-a, etc) [Blomqvist, 2005]. Figure.1 shows an example on the ontology. This example illustrates the components of the ontology as following:

- “*Person*” is a superclass;
- “*Student*” and “*Professor*” are subclasses of “*Person*” class;

- Also, “*Student*” and “*Professor*” are hyponyms of “*Person*” (their hypernym) where each of them has (is a) hyponym of “*Person*”.
- “*E-mail*” and “*Name*” are properties/attributes of the “*Person*” class;
- “*Student_no.*” and “*Research field*” are also properties/attributes of “*Student*” and “*Professor*” classes respectively;
- “*Is-a*” is a relation where (the student is a person) and (the professor also is a person).
- “*AI*”, “*Security*” and “*Big Data*” are individuals of the “*Research field*” class.

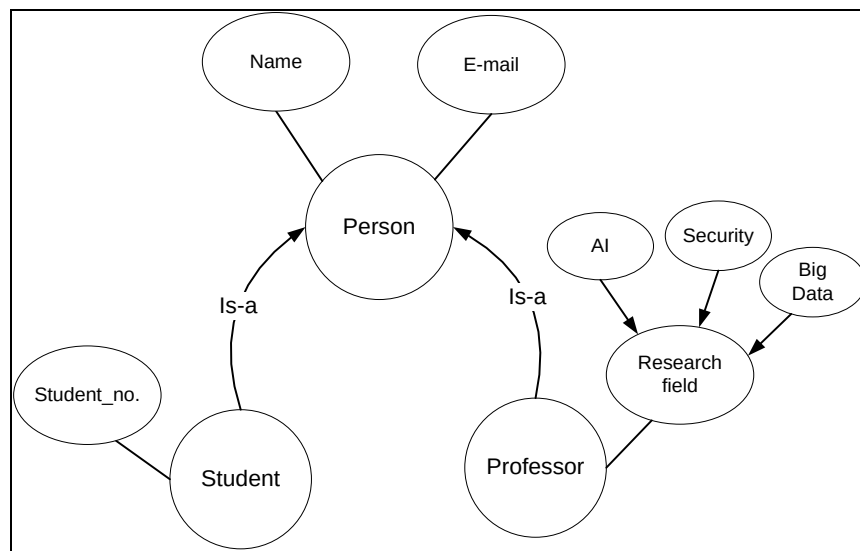


Figure 1. An example for the Ontology

The relations in ontology are classified as taxonomic or non-taxonomic relations as follows:

a) Taxonomic Relations:

Taxonomy is a hierarchical classification system: a tree that starts from a universal root concept and divides it into more specific child concepts. The taxonomy nodes correspond to concepts, connected by branches or edges directed from the root node towards the leaf nodes. The taxonomic relations organize the ontology by arrangement the entities in

its correct place. It provides the taxonomic layer of the ontology including parent/child or subClass/superclass [Nakashole et al. 2012]. For example, the blue is a color.

b) Non taxonomic relations:

Non taxonomic relations are complex relations between concepts that may have a single verb connecting two entities such as A worksFor B. This type of the relations detects how one concept can act upon another in the given domain. For example, a player can kick a ball [Cimiano & Völker, 2005].

Ontologies can be built either manually, semi-automatically or automatically. The main steps of the ontology construction are: (1) build glossary of terms such as concept, instance, relation and attribute, (2) build the taxonomic relations such as (Is-a hierarchy), (3) build diagrams for non-taxonomic relations, and (4) build concept dictionary [Corcho et al., 2005]. Figure 2 shows these steps.

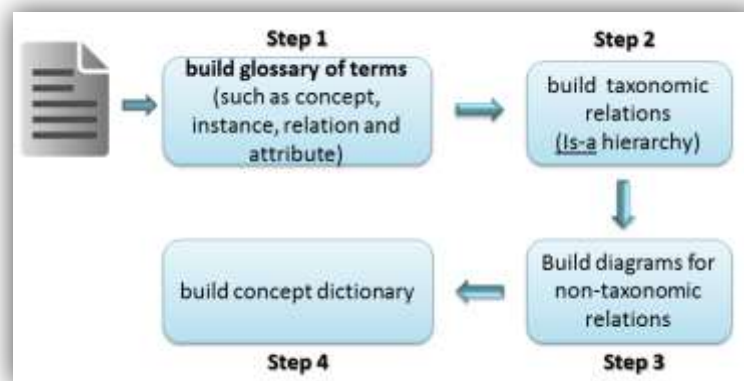


Figure 2: The main steps of the ontology construction process

There are a lot of tools that have been developed for building the ontology manually such as Apollo, OntoStudio, Protégé, Swoop, and TopBraid Composer. (Kapoor & Sharma) in 2010 presented a comparative study between these tools which showed that the “Protégé” is the best tool for building the ontology manually.

As we showed before that many researchers have tried to build the ontology semi-automatically or automatically because of the difficult of building it manually that called “Ontology Learning”.

2.2 Ontology Learning

Ontology learning can be defined as the set of techniques and methods used for building ontology semi-automatically or automatically from a given text corpus in which ontological elements such as concepts and relations are extracted automatically from different resources [Hazman et al. 2011]. It uses techniques and methods from different fields such as machine learning and natural-language processing [Al-Zoghby et al. 2018].

Because of ontologies importance in the semantic web, many research developed several ontology learning systems and approaches in different languages such as OntoLearn, Alvis, Text2Onto, and SPRAT [Gherasim et al., 2013]. Such systems aren't well for syntactically ambiguous languages such as Arabic compared to other languages such as English [Al Zamil & Al-Radaideh 2014]; where there are several challenges/problems face the process of Ontology learning when it is used with Arabic text. Such challenges are absence of capitalization, absence of diacritics, complicated morphology and the lack of resources. Specific terms in the Latin languages like English begin with capital letters such as proper names e.g. (“Ahmed”, “Mohammed”) and abbreviations e.g. (“ACM”, “IBM”). But this feature doesn't exist in the Arabic language because the Arabic language can't support the capitalization. Absence of this feature effects on the knowledge extraction task [Shaalán, 2014]. Arabic text also contains the diacritics  that “affect the phonetic representation and give different meaning to the same lexical form” [Shaalán, 2014]. Absence of the diacritics in the Arabic words leads to the ambiguity because different diacritics represent different meanings. Finding many different patterns for the one Arabic word is also from the characteristics of the Arabic language where each word can

consist of one or more prefixes, a stem or root, and one or more suffixes in different combinations, that lead to complicated morphology. The lack of the Arabic linguistic resources and tools that are designed according to the specific features of the Arabic language represents another challenge to the process of the semantic relation extraction and Ontology learning [Hawalrah, 2018]. These challenges effect on knowledge extraction from Arabic text. As a result, the Arabic language suffers a lack of Ontologies and the semantic web applications [Al-Arfaj & Al-Salman, 2014].

There is more than a method for evaluating the Ontology Learning (OL) approaches. One of them is computing the similarity between manually-built reference ontology, and an ontology that has been generated by applying a particular OL technique. It is assumed that a high similarity between the manually-modeled ontology and the OL-based ontology indicates a successful application of a particular OL technique. Also, the two straightforward evaluation measures (Precision and Recall) are used in evaluating this method [Albukhitan & Helmy]. The evaluation measurements such as “Precision, Recall, F-measure, and Accuracy” are the most used of the evaluation techniques [Al-Jumaily et al., 2012].

There are few efforts for constructing an automatic ontology for the Arabic language comparing with the other languages such as (English, Chinses). Extracting the terms and the semantic relations from the text are primary steps in buildings the ontology automatically so the following sections will review the previous works about term extraction, semantic relations extraction, and automatic ontology construction.

3 Term Extraction

Term extraction extracts the relevant phrases and terms from the text for a specific domain. It involves applying information extraction (IE) methods to extract terms. Named Entity Recognition (NER) is a subtask of the information extraction (IE) task that is used in recognizing the proper names in the text such as “Person”, “Location”, and “Organization”. There are three main approaches used in the NER namely: Rule Based NER, Machine Learning Based NER, and Hybrid Based NER. The rule based approach depends on the linguistic rules of the language. Some researchers used this approach for extracting the NER. The machine learning approach depends on the features of the named entity classes in a large training corpus. It overcome on some the Rule based challenges. The hybrid based approach combines between the two previous approaches to overcome some the drawbacks for each.

In the following, we review some of works on these approaches.

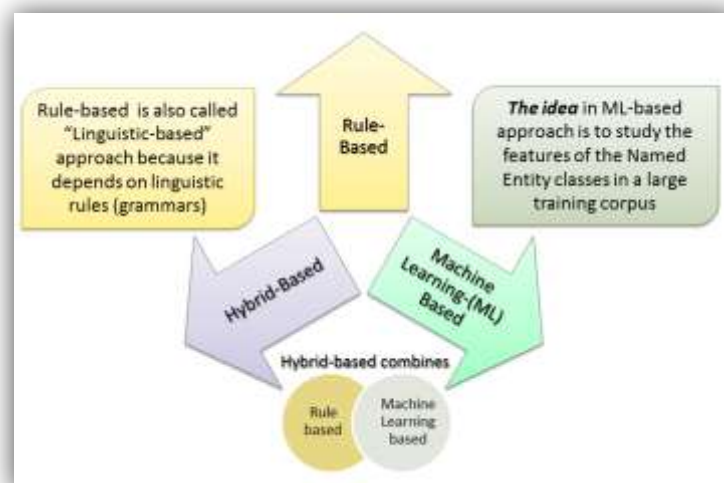


Figure 3 : The main approaches for Named Entity Recognition Task

3.1 Rule Based NER

Asharef et al. (2012) presented a rule-based approach to extract Arabic named entities from crime documents. To identify and classify named entities in Arabic crime text, several rules and patterns are applied. The approach includes three modules: (1) pre-processing module that contains these processes: sentence splitting, tokenization, and POS tagging; (2) second module is identifying the named entity and detecting of the tokens boundaries that belong to a named entity; and (3) Final module is the classification of the terms by using set of grammatical rules and patterns and gazetteer. The result showed that this approach is effective and the performance is satisfactory compared with the previous tasks on the crime domain. The drawbacks of this research are the needing to increase the tag set and the NEs labels and it is domain dependent. Another research was by **Btoush et al. (2016)** who used the rule based techniques to build a tool for the Name Entity Recognition for Arabic Language. The Named-Entity detector has applied rules on the text and has given the correct Labels for each word; the labels are for three named entities: Person, Location and Organization. The name entity detector was tested on a file that contains 490 words and successfully tagged 480 words.

The researches that used rule based approach play an important role in the process of named entities extraction. But this approach is domain dependent so, the researches achieved better results in specific domains only; so, if the same approach applied on another domain, the results will be different. Another problem of this approach requires an expensive manual effort and it is time-consuming; and used it with the languages that have complex morphology such Arabic, the problem increase.

3.2 Machine Learning Based NER

Alsayadi & ElKorany (2016) presented a new model depends on the machine learning approach. The aim of this model is to combine several linguistic features and to utilize syntactic dependencies to infer semantic relations between three named entities: person, organization and location. The proposed model helped overcome some of the orthographic and morphological problems of Arabic language. They used the conditional random field as machine learning classifier for recognizing the named entities. Experimental results show that this approach can achieve good performance where its F-measure was 87.86% for ANERCorp⁴ corpus; but they needed to test other ML algorithm to enhance the performance.

There are some researchers used the neural network algorithms for the named entity recognition task such as [Gridach, 2016]. **Gridach (2016)** proposed a novel neural network architecture by using combination of bidirectional Long Short-Term Memory (LSTM) and Conditional Random Field (CRF) for Dialectal Arabic and Modern Standard Arabic texts. Experimental results showed that the proposed model achieves state-of-the-art performance on publicly available benchmark for Arabic NER for social media.

Deep learning has been shown to perform significantly better than other approaches for different Natural Language Processing tasks including NER. **Helwe & Elbassuoni (2017)** presented a new approach called “deep co-learning” to detect and classify named entities in any Arabic text. The proposed approach used a small amount of labeled data to overcome on the problem of lacking the Arabic resources. The approach is based on a semi-supervised learning algorithm known as co-training. For training the proposed approach, they developed a Wikipedia article classifier using an LSTM deep neural network that was used to generate a semi-labeled dataset for the Arabic NER task. The experimental

⁴ Available for download from <http://users.dsic.upv.es/ybenajiba/>

results show that the proposed approach outperformed all other compared approaches.

3.3 Hybrid Based NER

Abdallah et al. (2012) presented an approach for integrating the rule-based approach with the Machine learning-based approach for Arabic named entity recognition. The authors focused on only three named entities (person name, location, and organization). For training and testing the proposed method, they used two annotated corpora: the ACE 2003⁵ Multilingual training set and ANERcorp corpus. Steps of building their rule-based system: (1) performing the recognition based on a dictionary lookup that containing lists of known named entities, and (2) using a parser, based on a set of grammar rules (represented as regular expressions). Steps of building the proposed integrated approach: (1) using the Stanford POS Tagger to compute some of the general features such as word category and affixation that are defined as machine learning features; (2) Complementing the rule-based features with the other extracted features; and (3) feeding all combining features to a decision tree classifier. The results proved that the proposed hybrid approach is better than the pure rule-based system. **Oudah & Shaalan (2017)** proposed another hybrid system that integrates both rule-based and machine learning-based NER approaches. Their proposed hybrid NER system is considered the state-of-the-art in Arabic NER according to its performance on standard evaluation datasets. They used ACE 2004⁶ Newswire standard dataset for extracting new rules for person, location and organization names recognition. They formulate each new rule based on two feature groups,

⁵ Available to BUID under License from <https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications>

⁶ (ACE) 2004 Newswire (NW): the NW data from the Arabic training dataset for the Automatic Content Extraction (ACE) evaluation conducted in 2004, which has been created by Linguistic Data Consortium (LDC)

(1) Gazetteers of each type of named entities and (2) Part-of-Speech tags. The results show that the proposed hybrid approach overcame the drawback of rule-based NER systems and it could improve the performance.

In some cases, the depending only on rule-based features doesn't improve the performance; and the depending only on machine learning based features doesn't improve performance. But when integrating the features of rule-based with Machine learning classifiers, in this case the performance can be improved.

Table 1: Summary of Works on Term Extraction from Arabic Text

Reference	Technique
<i>Asharef et al. (2012)</i>	Rule-Based
<i>Btoush et al. (2016)</i>	Rule-Based
<i>Alsayadi & ElKorany (2016)</i>	Machine Learning Based
<i>Gridach (2016)</i>	Machine Learning Based
<i>Helwe & Elbassuoni (2017)</i>	Machine Learning Based
<i>Abdallah et al. (2012)</i>	Hybrid Based
<i>Oudah & Shaalan (2017)</i>	Hybrid Based

4 Semantic Relation Extraction

Extracting the semantic relations from the text is an important step when building the ontology. Several researches proposed different methods for semantic relation extraction.

Hearst, (1992) is from early researchers who worked on the semantic relation extraction and his proposed algorithm considers the basic for many similar researches. He proposed an algorithm depend on the statistical- approach for the automatic extraction of the lexical relations like (is-a, kind-of or such as) by building lexical patterns of knowledge. The advantage of this algorithm is using

it as low-cost approach for automatic extraction of semantic lexical relations from text. But the main drawback is the inefficient performance on ambiguous text, such as Arabic text that leads some researchers such as [Al Zamil et al., 2014] to deal with the Arabic language to overcome such drawback. **Al Zamil et al., (2014)** proposed a technique for extracting semantic relations from Arabic text. The proposed technique is implemented as an enhanced version of Hearst's algorithm. The objectives of this research are: to extract semantic features of Arabic text, to propose relationships among concepts, and to propose a formal model of extracting ontological relations. The steps of this technique are as follows: (1) analyzing Arabic text using lexical semantic patterns of the Arabic language according to a set of features such as POS tag features; (2) building lexical syntactic patterns of Arabic text by enhancing the Hearst's algorithm; (3) expansion phase to avoid having redundant patterns that refer to the same concepts; and (4) finally, pattern filtering and aggregation. The advantage of this technique according to the results of this research, the technique is a good candidate for extracting ontological relations from Arabic text. But the results showed that the performance among different datasets is not systematic; where they used three different Arabic datasets, Holy Qur'an "Classical Arabic", newspapers "Modern Standard Arabic", and social blogs "unstructured Arabic texts". The Newspapers dataset has the highest performance compared with other datasets and the Blogs dataset has the lowest performance. The reason for this is the presence of few classification errors that had a negative impact on the performance of the proposed techniques. Table 2 shows the results of a comparison between the original Hearst's algorithm on Arabic texts from different datasets with the Al-Zamil technique on the same datasets which shows that Al-Zamil technique achieved the highest performance.

**Table 2: Performane Comparison between Al-Zamil technique and Hearst's algorithm
(Al Zamil et al., 2014)**

Average measurements	Dataset	Precision (%)	Recall (%)	F-Measure (%)
Original Hearst's Algorithm (Arabic Text)	Qur'an	46.74	50.32	48.48
	Newspaper	51.23	61.35	55.84
	Blogs	47.43	53.45	50.26
Enhanced Version of Hearst's Algorithm (Arabic Text)	Qur'an	76.28	82.94	79.47
	Newspaper	89.77	84.49	87.05
	Blogs	69.66	74.70	72.09

5 Automatic Ontology Construction

Building the ontology automatically can be by using different techniques: NLP techniques like tokenization, stemming and splitting; Machine Learning techniques and Deep learning techniques. There are several research proposed systems and approaches presented to build the ontology automatically for different languages such English and Arabic languages. The following are some previous works of such systems for Arabic language:

Al-Rajebah, & Al-Khalifa (2014) proposed a system which considered each article's title as a concept and extracted its semantic relations from infoboxes and the list of categories contained in each Wikipedia article. The system consisted of three main phases, (1) filtration, (2) extraction and (3) ontology generation. For each article, the infobox was extracted from the articles text. Each infobox was then parsed to extract (hasFeature), (isRelatedTo) and (hasCategory) relations. The (hasFeature) relation, defined articles features and their values, the (isRelatedTo) relation identified the related Wikipedia articles, and the (hasCategory) relation extracted article's categories. Then, the final Ontology was generated and written as an OWL file. The evaluation result showed that the average precision of the system equals to 65%. This performance was because of the presence of more duplication in the concepts and relations.

Harrag et al. (2014) proposed a new representation tool for the Quran. A linguistic pattern-based approach is exploited to extract specific concepts from the Quran, while the conceptual relations are found based on association rules technique. The authors proposed the tool without any evaluation are presented but they predicted that the proposed Quran ontology will offer a new and powerful representation of Quran knowledge, and the association rules will help to represent the relations between all classes of connected concepts in the Quran ontology.

Al-Arfaj and Al-Salman, (2014) proposed a framework for Arabic ontology construction based on Hadith texts (sayings of Prophet Mohammed). Figure 4 shows the proposed framework. It consists of the following phases: (1) pre-processing of corpus, (2) extracting the concepts, (3) extracting the relations between concepts, and finally (4) building the ontology. The authors discussed the challenges of constructing ontology from Arabic texts and the solutions for each but there aren't results to show.

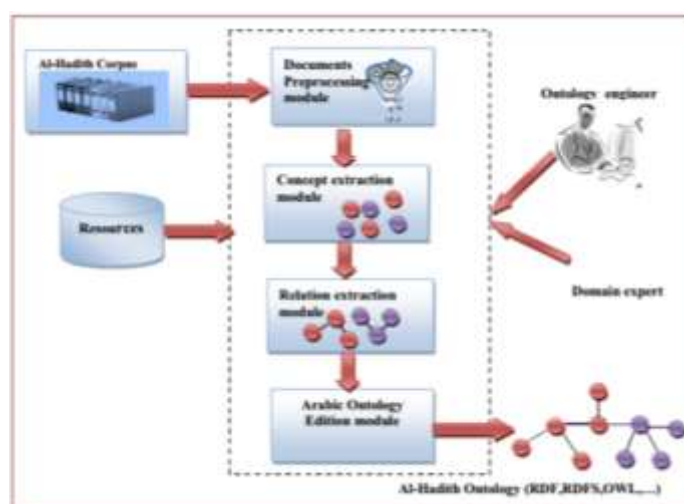


Figure 4: A framework for Ontology construction from Arabic texts
[Al-Arfaj and Al-Salman, (2014)]

Hawalah, (2018) proposed another approach for building an Arabic ontology from multiple resources (publicly available directory, rich data from the Internet, and Arabic online directory). The proposed approach consisted of two main phases: (1) building and extracting an Arabic ontology from a publicly available directory, (2) enhancing the Arabic ontology by using richer information from the Internet. The results showed that the proposed approach provides more accurate classification results than the traditional classification algorithm. But the drawback of this research is using small number of items in the test dataset that leads to unreliable results.

Albukhitan & Helmy (2013) presented a development of an Arabic semantic annotation tool with integrated information extraction for the domains of food, nutrition and health. The proposed method aimed to developing Arabic OWL ontologies related to those domains then these ontologies are integrated together to produce one ontology. The proposed method used linguistic patterns to discover relevant relationships between the named entities in the Arabic Web resources. The extracted information was then associated to the corresponding concepts and object properties of the developed ontology to produce the RDF metadata for the corresponding Web resources. Empirical evaluations of the proposed method show promising precision and recall.

Albukhitan & Helmy (2016) presented an Ontology Learning framework based on some available NLP tools for Arabic text, utilizing the GATE text analysis system for corpus and annotation management. The main steps in this framework were as following: (1) Text Preprocessing using NLP techniques, (2) Concept Recognition using Data Mining techniques and algorithms based on the statistical measures, and (3) Relation extraction using the traditional techniques in addition to some algorithms are built to enhance the accuracy. To evaluate the effectiveness of the proposed framework, a set 100 Arabic documents were

selected and manually annotated by hand. Manual annotation was limited to concepts and relationships related to document domains. The overall system performance for the Precision was promising while the recall was low.

The deep learning techniques also shared in solving the ontology learning problem for the Arabic language texts. *Albukhitan et al. (2017)* proposed an Arabic ontology learning system using deep learning. The authors investigated the performance of implementing deep learning with Arabic ontology learning tasks using major models such as Continuous Bag of Words (CBOW) and Skip-gram. The proposed system consists of five steps namely: Data Acquisition, Preprocessing, and Deep Learning, Ontology creation, and ontology alignment. The experimental results showed that the proposed system is better than the traditional ontology learning systems. Table 3 shows a comparison between the work of (Albukhitan & Helmy, 2016) as Traditional Ontology Learning (OL) and (Albukhitan et al. 2017) as Ontology Learning using Deep Learning approach. It shows that using the Deep Learning approach can enhance the performance.

Table 3: Comparison between the work of (Albukhitan & Helmy, 2016) and (Albukhitan et al. 2017) [Albukhitan et al. 2017]

Approach	Task	Count	Extracted	Correct	Precision	Recall
Traditional OL Approach	Concepts	5022	4522	3261	72.1%	64.9%
	Taxonomy Relation	650	551	461	83.7%	70.9%
	Non Taxonomy Relation	180	117	88	75.2%	48.9%
OL using Deep Learning Approach	Concepts	5022	4803	3861	80.39%	76.88%
	Taxonomy Relation	650	591	494	83.59%	76.00%
	Non Taxonomy Relation	180	122	93	76.23%	51.2%

Table 4 shows a summary of some researches on Automatic Ontology Learning from Arabic Text.

6 Comments and Limitations of Research Work on the automatic Ontology construction

From the previous work on the automatic Ontology construction, we concluded the following:

- There are few efforts for building Arabic ontology automatically.
- Most of them focused on specific domain.
- Using small dataset can effect negatively on results.
- Using the Hybrid approaches (Linguistic and Statistical) give better results.
- Using the deep learning techniques such as: Continuous Bag of Words (CBOW), Skip-Grams, and Word embedding can improve the ontology learning process.
- The type of Arabic language like (classical Arabic, Modern Standard Arabic, or Dialect Arabic) effects on the performance.
- The majority of the work dedicated to extract the semantic relationships was for the purpose of building Ontologies.

Table 4: Summary of some Works on Automatic Ontology Learning from Arabic Text

Ref.	Techniques	Datasets	Results		Description/Limitations
<i>Al-Rajebah, & Al-Khalifa (2014)</i>	Linguistic Techniques (NLP tools + pattern matching)	Arabic Wikipedia Articles	Precision = 65 %		More concepts and relations are duplicated.
<i>Harrag et al. (2014)</i>	Linguistic + statistic (Association Rules)	12 Surahs from the Holy Quran related to the stories of prophets in the Qur'an 1407 verses, 16153 words.	No Result to show		They presented a new representation tool for Quran.
<i>Al-Arfaj & Al-Salman, (2014)</i>	Linguistic + statistic + Data Mining techniques	Hadith corpus	No Result to show		They proposed their framework depending on evaluating each step by an expert.
<i>Albukhitan & Helmy (2013)</i>	linguistic techniques (NLP Tools) + Statistical techniques	5000 Web documents about food, nutrition and health	More relationships are for the assessment such: Nutrition - Body Functions: Precision = 69% Food - Body Functions: Precision = 80%		The results of some relationships aren't satisfied.
<i>Albukhitan & Helmy (2016)</i>	NLP + statistical and data mining techniques	100 Arabic documents	<u>Concepts:</u> Precision = 84% Recall = 77% <u>Taxonomic Relations:</u> Precision = 67% Recall = 46%	<u>Non Taxonomic Relations:</u> Precision = 73% Recall = 29%	The results of some relationships are low recall
<i>Albukhitan et al. (2017)</i>	NLP + Deep Learning	About 5 thousand words	<u>Concepts:</u> Precision = 80% Recall = 77% <u>Taxonomic Relations:</u> Precision = 83% Recall = 76%	<u>Non Taxonomic Relations:</u> Precision = 76% Recall = 51%	The results of some relationships are low recall

7 Conclusion

After years of development, Ontology Learning Process (OL) for Latin-character languages, such as English, has been refined greatly. Arabic, however, possesses several challenges that make OL more difficult. In This paper, we review these challenges and survey some of the recent research aiming to improve the Arabic Ontology Learning. We found that most of the research works focused on three main issues: (1) extracting the Terms, (2) extracting the semantic relations, and (3) building the ontology automatically. “Extracting the terms” is through three approaches: rule based approach, machine learning approach and hybrid approach. The survey showed that the hybrid approach is the best technique for extracting the terms from the text. The works of “extracting the Semantic relations” process showed that using the hybrid techniques such the NLP techniques with (the statistical techniques or Data Mining techniques) improve the process of semantic relation extraction from the Arabic text. From the review that presented in this paper, we concluded that using the Deep Learning techniques will help in improving the Ontology Learning process.

In the future, we will present a new approach for ontology learning process that will help in solving the problems that are found out in the previous works, and trying to improve the performance. The different techniques such as Natural Language Processing, Machine Learning, Deep Learning and Data mining techniques will be investigated to choose the best among them to be used for each step in developing the approach.

References

- Abdallah, S., Shaalan, K., & Shoaib, M. (2012). Integrating rule-based system with classification for Arabic named entity recognition. In Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Springer- Verlag, Berlin Heidelberg, 311-322.
- Al Zamil, M. G. & Al-Radaideh, Q. (2014). Automatic extraction of ontological relations from Arabic text. Journal of King Saud University-Computer and Information Sciences, 26(4), 462-472.
- Al-Arfaj, A., & Al-Salman, A. (2014). Towards ontology construction from Arabic texts-a proposed framework. In Computer and Information Technology (CIT), 2014 IEEE International Conference on (737-742). IEEE.
- Albukhitan, S. & Helmy, T., (2016). Arabic Ontology Learning from Unstructured Text. IEEE/WIC/ACM International Conference on Web Intelligence (WI). 492-496. DOI: 10.1109/WI.2016.0082.

- Albukhitan, S., & Helmy, T. (2013). Automatic ontology-based annotation of food, nutrition and health Arabic web content. *Procedia Computer Science*, 19, 461-469.
- Albukhitan, S., Helmy, T., & Alnazer, A. (2017). Arabic ontology learning using deep learning. In *Proceedings of the International Conference on Web Intelligence* (1138-1142). ACM.
- Al-Jumaily, H., Martínez, P., Martínez-Fernández, J. L., & Van der Goot, E. (2012). A real time Named Entity Recognition system for Arabic text mining. *Language Resources and Evaluation*, 46(4), 543-563
- Al-Rajebah, N. I., & Al-Khalifa, H. S. (2014). Extracting ontologies from Arabic Wikipedia: A linguistic approach. *Arabian journal for Science and Engineering*, 39(4), 2749-2771.
- Alsayadi, H. A., & ElKorany, A. M. (2016). Integrating semantic features for enhancing arabic named entity recognition. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 7(3), 2016.
- Al-Zoghby, A. M., Elshiwi, A., & Atwan, A. (2018). Semantic Relations Extraction and Ontology Learning from Arabic Texts—A Survey. In *Intelligent Natural Language Processing: Trends and Applications* (199-225). Springer, Cham.
- Asharef, M., Omar, N., Albared, M., Minhui, Z., Weiming, W., Jingjing, Z., . . . Yu, F. (2012). Arabic Named Entity Recognition In Crime Documents. *Journal of Theoretical and Applied Information Technology*, 44(1), 1-6.
- Azevedo, R., Freitas, F., Rocha, R. G., Alves de Menezes, J. A., de Oliveira Rodrigues, C. M., & Silva, G. D. F. (2014). An Approach for Learning and Construction of Expressive Ontology from Text in Natural Language. Paper presented at the Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on.
- Barforush, A. (2012). Ontology learning: revisted. *J. Web Eng.*11 (4), 269–289
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities [Electronic resource]. *Scientific American*.
- Bhaskar K. & Savita S. (2010). A Comparative Study Ontology Building Tools for Semantic Web Applications. (IJWesT) Vol.1, Num.3
- Blomqvist, E. (2005). Fully Automatic Construction of Enterprise Ontologies using Design Patterns: Initial method and first experiences On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE (1314-1329). Italy: Springer.

- Btoush, M. H., Alarabeyyat, A., & Olab, I. (2016). Rule based approach for Arabic part of speech tagging and name entity recognition. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 7(6), 331-335.
- Cimiano, P., & Völker, J. (2005). *Text2Onto Natural Language Processing and Information Systems* (227-238), Berlin Heidelberg : Springer.
- Corcho, O., Fernández-López, M., Gómez-Pérez, A., & López-Cima, A. (2005). Building legal ontologies with METHONTOLOGY and WebODE. In *Law and the semantic web* (142-157). Springer, Berlin, Heidelberg.
- Gantayat, N. (2011). *Automated Construction of Domain Ontologies from Lecture Notes*. (Unpublished Master Thesis), Indian Institute of Technology, Bombay, Indian, Bombay.
- Gherasim, T., Harzallah, M., Berio, G., & Kuntz, P. (2013). Methods and tools for automatic construction of ontologies from textual resources: A framework for comparison and its application. In *Advances in Knowledge Discovery and Management* (177-201). Springer, Berlin, Heidelberg.
- Gridach, M. (2016). Character-aware neural networks for Arabic named entity recognition for social media. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)* (23-32).
- Gruber, T. (1993). What is an Ontology. WWW Site <http://www-ksl.stanford.edu/kst/whatis-an-ontology.html> (last visit 10 – 7 - 2018).
- Harrag, F., Al-Nasser, A., Al-Musnad, A., Al-Shaya, R., & Al-Salman, A. S. (2014). Using association rules for ontology extraction from a Quran corpus. In *Proc. 5th Int. Conf. Arabic Language Process* (1-8).
- Hawalah, A. (2018). A Framework for Building an Arabic Multi-disciplinary Ontology from Multiple Resources. *Cognitive Computation*, 10(1), 156-164.
- Hazman, M., El-Beltagy, S. R., & Rafea, A. (2011). A survey of ontology learning approaches. *database*, 7(6).
- Hearst, M.A. (1992). Automatic acquisition of hyponyms from large text corpora. In: *Proceedings of the 14th conference on Computational linguistics*, vol. 2, (539–545). Association for Computational Linguistics.
- Helwe, C., & Elbassuoni, S. (2017). *Arabic named entity recognition via deep co-learning* (Doctoral dissertation).
- Nakashole, N., Weikum, G., & Suchanek, F. (2012). *PATTY: A Taxonomy of Relational Patterns with Semantic Types*. Paper presented at the *Proceedings of*

the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Oudah, M., & Shaalan, K. (2017). NERA 2.0: Improving coverage and performance of rule-based named entity recognition for Arabic. *Natural Language Engineering*, 23(3), 441-472.

Petrucci, G., Ghidini, C., & Rospocher, M. (2016). Ontology learning in the deep. In *European Knowledge Acquisition Workshop* (480-495). Springer, Cham.

Shaalan, K. (2014). A survey of Arabic named entity recognition and classification. *Computational Linguistics*, 40(2), 469-510.

Somodevilla, M. J., Vilariño Ayala, D., & Pineda, I. (2018). An Overview on Ontology Learning Tasks. *Computación y Sistemas*, 22(1).

Vortex swarm optimization

Ahmed sabry A.elrahman *

Hesham A. Hefny †

Abstract

this paper proposed a new swarm intelligence algorithm called vortex swarm optimization, Inspired from vortex collective behavior where individuals move around a common known center .The algorithm performs the two basics functions that any searching algorithm had (exploitation and exploration) by cycling, changing vortex direction, increase propulsion force, Algorithm skip local minimum and by reduce interaction range the algorithm changes from exploration to exploitation. The algorithm was tested over five benchmark functions and shows that it a competitive with other powerful algorithms.

Key Words: vortex swarm optimization, collective behavior, vortex collective behavior, optimization

1 Introduction

Swarm intelligence has been attracted many researchers in last decades where researchers adopts intelligence mechanisms developed by organisms in nature. One of them is collective behavior where interactions of many individuals lead to movement of whole group, There are different major types of collective behavior:”swarms”-where there is weak movement synchronization and the group is almost align;”polarized”-where there is high movement synchronization and the group move direct with high alignment;”vortex”- where individuals move around empty core center, Bat doughnut, duck swirl, fish torus are different names for collective vortex behavior. This type of collective behavior occurs at different size levels from pair to millions of individuals, Vortex increases efficiency of feeding, save group from predators, improve decision-making efficiency. Here we will focus on this type of collective behavior (vortex), And Develop a new swarm intelligence algorithm called vortex swarm optimization to solve optimization problems.

2 Collective Behavior

Schools of fish, bird’s flocks, and insect’s swarms are considered as collective behavior where a group of individuals have to make quick decisions about what actions they will perform in dangerous and uncertain environment. Note that all actions is done synchronously [14] [15]. Most of people think that collective behavior considered to only animal movements but it also considered to bacteria, insects, birds, atoms and molecules. Here we will ask an important question, what is collective behavior? Collective behavior happen in systems that consists of many units interacting with each others, the interactions may be simple or complex and occur between the unit and its neighbors so the unit behavior completely determine by the collective effects of its neighbors so the behavior of a lonely unit will be completely different from a unit with one or more

* Department of computer science, Institute of Statistical Studies and Research, Cairo University.

†Prof. Computer Science, Department of computer science, Institute of Statistical Studies and Research, Cairo University.

neighbors [9]. There are multiple types of collective behaviors [17] one of them is called “polarized” where particles (fish, birds,...) moving together straight ahead to a specific destination, Another type is called vortex collective behavior where particles circle coherently .

3 biological basis (why collective behavior?)

Why do animals flock together? There are some reasons for particles to perform collective behavior, the first reason is to protect themselves from predator by many-eyes theory where individuals are everywhere as many eyes is better than one so any individual feels any surrounding predator it will exchange that information with others individuals to escape from it ,and if the predator attack the group the individual probability to be eaten will be reduced compare to the probability if it alone ,the second benefit of being in a group is foraging because searching for food will be easy due to many information food resources provided by many individuals in group whose know the locations of food, the third benefit is finding mate as more the group having individuals the more chances the individual have to find mate, the fourth benefit is energy savings as individuals reduce their energy required to move by putting them- selves behind others. Finally be in a group is very important in migration as the individual needs some navigational informations provided by others [6].

4 CAUSES OF VORTEX FORMATIONS

the easy way for collective behavior is to mimic your neighbors direction to go to group destination (food source, safety place,...), But why organisms adopt vortex behavior, or in other word why I go to the same place that I was before and repeat this behavior again and again? What are the reasons for adopting vortex behavior? Are organisms do vortex collective behavior by their will?, here in this section we will insulate the different reasons for vortex collective behavior.

4.1 Maladaptive Examples

Maladaptive behaviors refer to animal’s disability to adjust to particular situations. These behaviors reflect a normal animal in an abnormal environment. Caterpillars follow each other and if the first individual follows the last one then a vortex will form. And if any individual changes its circular motion this individual will make this after several days by stopping following the individual in the front due to starvation so they break the vortex ”Fig.1 (A)”. Ant mill [22] is a phenomenon where an army ants group are separated from their colony and lose their tracking pheromone and begin to follow one another, forming a circular shape ,Army ants move until death because of moving in circular motion, Even if they do not die they suffer from energy and time losing ”Fig.1 (B)”.

4.2 Feeding Function and Access to Other Resources

Vortex movements in some cases lead to benefits for all individuals'. When filter-feeding species adopt vortex motion they releasing additional nutrients in the water. When shoveler ducks [18] adopt circular motion they cause nutrients to rise up to water surface so they can filter it from water "Fig.2 (A)". Some species adopt vortex behavior to ensure that all group individuals have access to resources location (e.g., best temperature, light, oxygen). Moreover vortex behavior change the distribution of the resources to reduce concentration differences of food and oxygen or temperature across locations.



Fig. 1. (A) vortex of Caterpillars [17] (B) Ant mill [17]

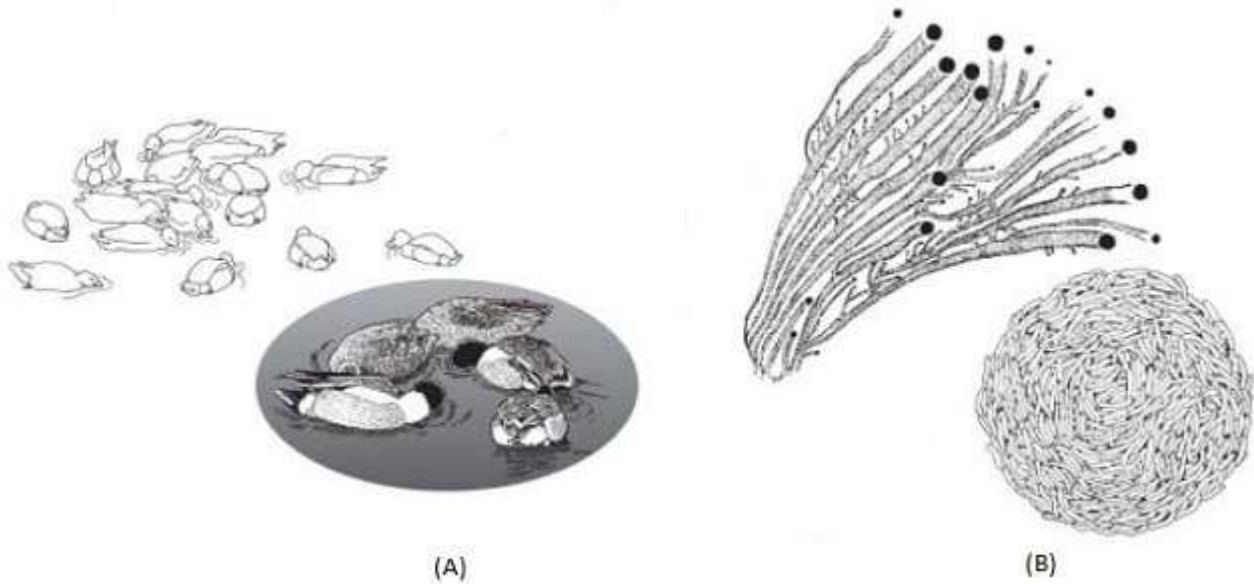


Fig. 2. (A) vortex of shoveler ducks [17] (B) Paenibacillus vortex Bacteria [17].

4.3 Locomotory Function

p.vortex bacteria [19] adopt circular motion to facilitate locomotion for high densities individuals by producing a wetting fluid on surface so the vortex moves as a one unit. Each cell carry small amount of fluid, the cell releases it when it feels that the surface is dry, and when cell go to a place where the fluid is below threshold value (vortex boundary) its motion will be reflected as there is a boundary. The propulsion force varies from cell to other based on the concentration of chemo-attractant. When the chemo-attractant is high the force is decreased and that leads the cells to move towards the direction of chemo-attractant. And when the chemo-repellent increases the propulsion force will be increased

4.4 Anti-predator Function

When fishes is attacked by a predator they come together and form vortex shape for many reasons : first to make themselves as one big animal to scare the predator and make it think twice before attacking them ,second vortex collective behavior is useful for organisms that had large blind spot in their visual field as vortex reduce the time required for predator direction detection and reduce the time required to alarm others as individuals everywhere and they are in all directions so any individual see a predator will alarm others to move vortex in the opposite direction, it has been shown that vortex collective behavior is the best type of all collective behavior types that reduce the risk of predatory attacks. [17].

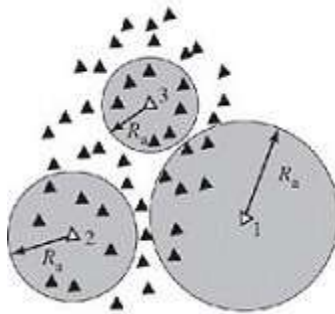


Fig. 3. interaction range is modified according to individual needs

4.5 Collective Decision-Making and Information Transmission

Collective decision-making and information transfer is another benefit for collective vortex behavior as the entire group remains in the same place because there is no consensus on the direction that the group will move to. If sufficient number follow an individual it becomes a temporary group leader and then the vortex will be broken (e.g. Bat doughnut).

5 Vortex Scales

to develop algorithm based on vortex collective behavior we have to know vortex scales, like how many individuals they should have to form vortex shape?, How long the vortex will last? , which direction vortex will select? What is the perfect sensory range individuals should have to

form vortex? All these questions we have to answer to know how to simulate vortex collective behavior.

5.1 Number of individuals

How many individuals must vortex have? ,some organisms do vortex behavior by only two individuals (Northern shoveler ducks) while bacteria vortex(p.vortex) includes millions of cells, so the number of individuals vortex must have depending on the nature of the organisms that perform vortex behavior, Perhaps it related to the body size of the individuals.

5.2 Interaction range

Every individual have sensory range [16] [24] in which it can exchange information with others that allows change in direction to be amplified and that make a wave of turning in all individuals in the group. This range influence and be influenced by others ,By this interaction range every individual have very big sensory which is the summation of sensory ranges of all group, That make "effective range" which perceive much larger than its sensory range. It is hard to remain the group cohesive if individual respond to its close neighbors so Individuals may modify their interaction range and increase it to make the group remain cohesive to interact with more neighbors, and if the interaction range full of neighbors, individual decrease interaction range to avoid collisions and to avoid individuals from being all in one small space. Experiments show that [21] the perfect interaction range to form vortex is equal to vortex radius.

5.3 range of time

How long vortex movements last is depending on the different mechanism that leads to vortex movement, Golden shiners (*Notemigonus crysoleucas*) move in vortex shape for second or minutes. In contrast some vortices last until a perturbation stops it [17].

5.4 Vortex Direction

Both clockwise and anticlockwise was observed in all known vortex behavior where almost individuals move in the same direction and when multiple vortex occur in the same time both directions can be adopted but only one direction the separate vortex can adopt [17] [12].

6 Mechanisms

There are many mechanisms that organisms use to form vortex shape here we will insulate these different mechanisms.

6.1 Nonsocial Attractive Force

The first mechanism that generate vortex by concentrate individuals in small area (Attract to a point) with two conditions: first all individuals moving with nonzero speed, second, two individuals

Cannot be in the same place at the same time in other word they have to avoid collisions. In this mechanism there is not any attractive forces between individuals .the vortex centered on the attractive point and last until attractive stimulus disappears [17] [18] with motion given by:

$$x_i(t + \Delta t) = x_i(t) + v_i(t + \Delta t)\Delta t \quad (1)$$

Update the position x_i and velocity v_i at fixed time step of length $\Delta t = 0.1$

$$v_i(t + \Delta t) = v_i(t) + F_i^{all}(t)\Delta t \quad (2)$$

Where $F_i^{all}(t)$ represents the force action on individual (i) at time (t)

$$F_i^{all}(t) = F_i^{drag}(t) + F_i^{avoid}(t) + F_i^{food}(t) + F_i^{stoch}(t) \quad (3)$$

The force $F_i^{drag}(t)$ represents the friction in system, $F_i^{avoid}(t)$ represents collision avoidance with others in interaction range (r), $F_i^{food}(t)$ represents attraction to highest nutrient in sensory range (r), $F_i^{stoch}(t)$ represents stochastic force.

$$F_i^{drag}(t) = -\gamma v_i(t) \quad (4)$$

The factor $-\gamma$ represents the friction factor

$$F_i^{avoid}(t) = \sum_{j \neq i} \frac{\theta(r - d_{ij})(x_i(t) - x_j(t))}{|x_i(t) - x_j(t)|} \quad (5)$$

Where d_{ij} is the Euclidean distance between individuals i and j and the Heaviside function θ implements the active interaction range (r).

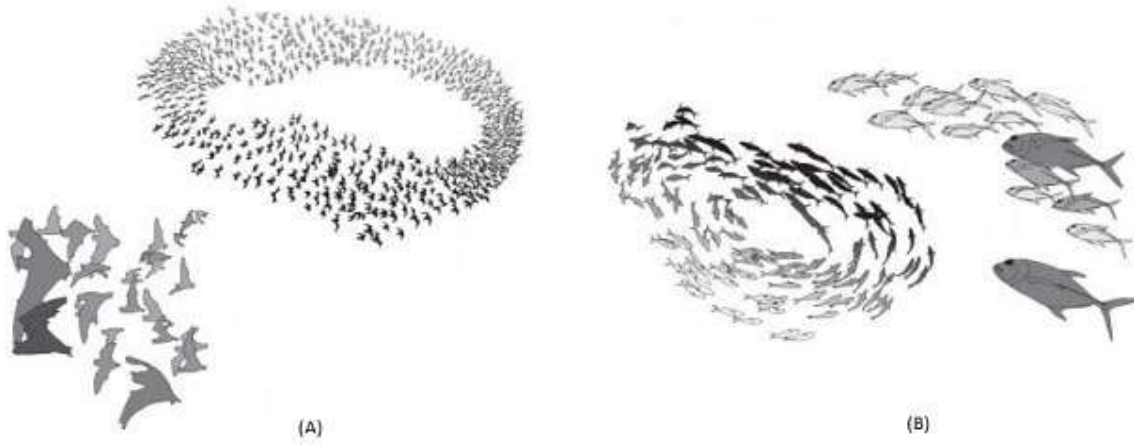


Fig. 4. (A) Bat doughnut [17] (B) fish bait ball [17]

6.2 Nonsocial Repulsive Force

Repulsion can make similar effect to attraction, repulsion from constraints surrounding group like walls or obstacles or predators around the group, if individuals move with nonzero speed and with collision avoidance [19].

$$\frac{d\vec{v}_i}{dt} = \mu[(\vec{v}_{i,\varepsilon}) - \vec{v}_i] - v\vec{v}_i + F\frac{\vec{v}_i}{v_i} - k\nabla\rho + \vec{\varepsilon} \quad (6)$$

Where μ , v , F , k coefficients are parameters of the model $\mu[(\vec{v}_{i,\varepsilon})]$ denotes the average direction of neighborhood individuals in radius (ε) of the i th particle, $v\vec{v}_i$ represents friction force, $F\frac{\vec{v}_i}{v_i}$ donates to propulsion force of the particle, ρ represents local density (number of particles in a unit area), and $\vec{\varepsilon}$ represents stochastic force. Note that there is a circular wall around the individuals play as repulsion force.

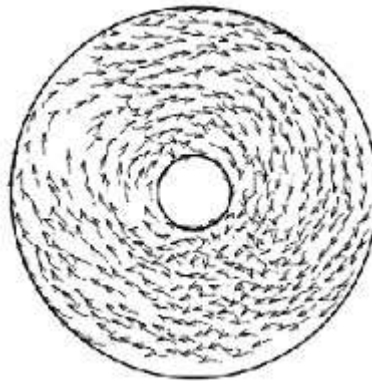


Fig. 5. circular motion with repulsion circular boundary [19].

6.3 Stigmergy

Stigmergy is a modification in the environment by individual behavior that affects other individual's behavior lead to collective behavior (ant mill) [17] [18].

6.4 Social Forces

Iain D. Couzin [1] proposed a three-dimensional self-organizing model of group formation where N individuals ($i=1,2,...,N$), c_i is position vector, v_i is unit direction vector and t is discrete time steps with regular spacing $\pi = 0.1s$. In each time step individuals determine their position and orientation by three non-overlapping behavioral zones.

First: Individuals keep minimum distance between themselves and others all the time, this rule is the most important rule observed in animals to avoid collision. When individual(s) n_r are get too close to individual i at time t (be in zone of repulsion (ZOR)), individual i will move away from its neighbors n_r within this zone"eq7"

$$d_o(t + \pi) = \sum_{j \neq i}^{n_r} \frac{r_{ij}(t)}{|r_{ij}(t)|} \quad (7)$$

Where $r_{ij} = (c_j - c_i)/|(c_j - c_i)|$ is a unit vector in the direction of the neighbor j.

second: Individual performing alignment with local neighbors ($n_o > 0$) if they are in zone of alignment (zoa) "eq8"

$$d_o(t + \pi) = \sum_{j \neq i}^{n_o} \frac{v_j(t)}{|v_j(t)|} \quad (8)$$

third: Individual performing attraction to other neighbors n_a if they in zone of attraction (zoa) where the distances between itself and other neighbors is big "eq9 "

$$d_a(t + \pi) = \sum_{j \neq i}^{n_a} \frac{r_{ji}(t)}{|r_{ji}(t)|} \quad (9)$$

The attraction rule performs by individual to avoid being alone while the orientation minimizing number of collision between individual and others. This mechanism form vortex by direct interactions between individuals who have short-live visual, long relatively range of attraction (zoa) and short relatively orientation range(zoo) which mean that group cannot all align so group individuals circling.

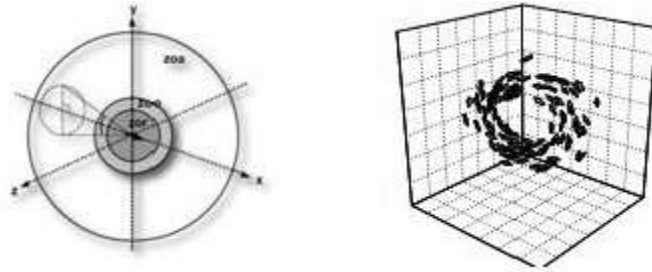


Fig. 6. Three-Circle model [1]

7 vortex swarm optimization

vortex swarm optimization play two roles foraging and feeding (exploration and exploitation) .foraging behavior is a behavior where individual interaction range is static and the propulsion force is random relativity big value . the individuals cover almost directions around the common center and there is a mutation in one dimension for random individuals to make the vortex changes its direction to make the individuals cover all directions around the common center(global best location) so when algorithm find a better solution around the best location the individuals will swim in this direction and begin to repeat this step until either reach the maximum steps or there are not any better solution than the center, so the algorithm will reach the best promising areas and vortex will overcome any local minimum .After while algorithm will play another role, feeding behavior(exploitation) where interaction range is decreasing every iteration and propulsion force is decreasing also so all individuals will get close to the best promising area and enhance solution by searching better solution around it.

Algorithm Phases: this algorithm contains two basic phases:-

1. cycling (form vortex) "Fig.7"
2. swimming (move towards food position) "Fig.8"

Algorithm Behaviors: this algorithm has the two basic phases in all of its behaviors but with some changes in radius and propulsion force

1. foraging Behavior: Where individuals form vortex with static radius value and relativity big random propulsion force value to search for food and when the individuals know the direction of food(as they cover all directions) they swim in that direction ,the static radius and random propulsion force save the algorithm from stuck in local minimum "Fig.8".

2. feeding Behavior: Where individuals form small vortex with decreasing radius value and decreasing propulsion force value to enhance current best solution "Fig.9".

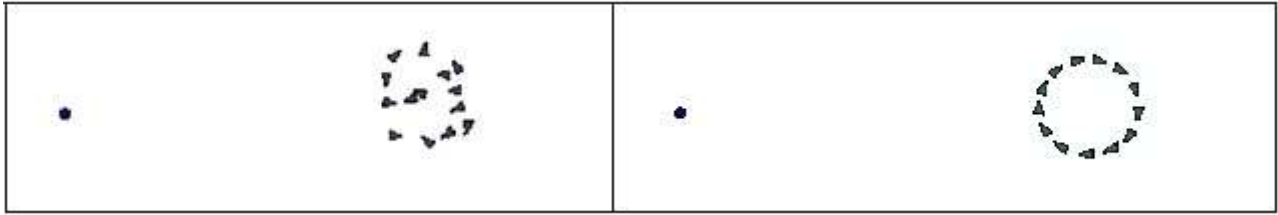


Fig. 7. Cycling phase :(A)At initial time swarms try to form vortex (B)vortex shape is formed

7.1 The model

to formulate phases and behavior of the algorithm we have to formulate cycling , swimming phases, foraging and feeding behaviors.

1. Cycling: in this phase algorithm form vortex and based on the previous mechanisms algorithm can form it by making all individuals follow three rules

Separation: which refers to collision avoidance of the individuals, with this rule any two? particles cannot be in the same place at the same time"eq5", "eq7".

Alignment: which refer to velocity matching with neighbors in sensory range"eq6" "eq8"

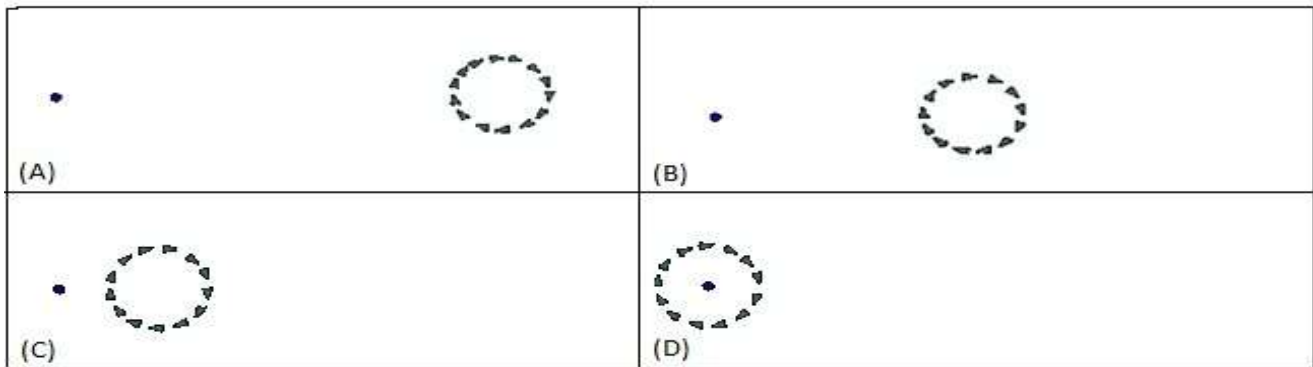


Fig. 8. Swimming phase with foraging behavior:(A-D) vortex detect the food position and go in its direction until vortex reach to it

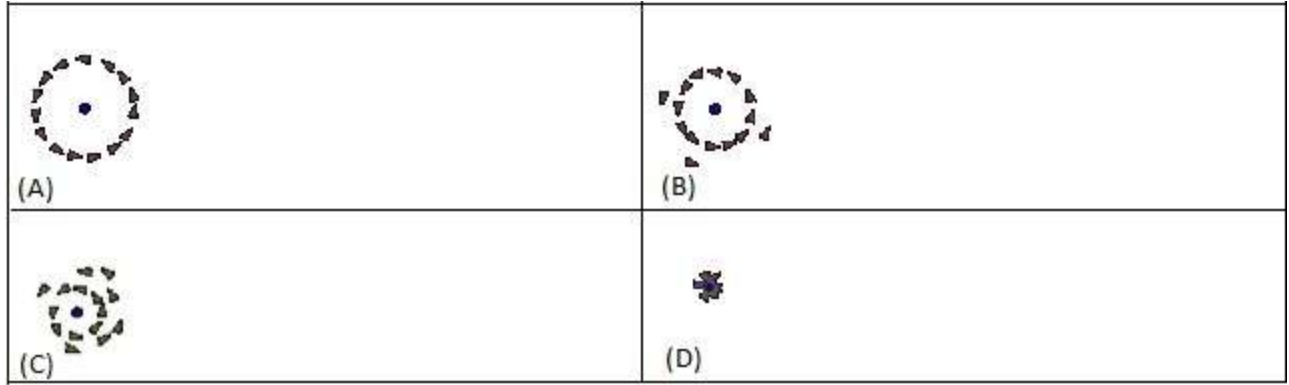


Fig. 9. Feeding behavior :(A-D) vortex detect the food position and get close to its position

Cohesion: either by reflecting circular boundary or by attracted to specific point”eq3” ”eq9”.

All these factors mathematically modeled as following: separation is calculated as follows [7] [20]:

$$S_i = - \sum_{j=1}^{n_r} \vec{x}_i - \vec{x}_j \quad (10)$$

Where $\vec{x}_i$ is the position of the i th individual j is the position of the j th individual, n_r is number of neighbors in i th sensory range with radius (r).

Alignment is calculated as follows:

$$A_i = \frac{1}{n_r} * \sum_{j=1}^{n_r} \vec{v}_j \quad (11)$$

where $\vec{v}_j$ is the j th neighbor’s velocity , n_r is number of neighbors in the sensory range radius (r) of particle i

The cohesion rule formulated as follows:

$$C_i = C - \vec{X}_i \quad (12)$$

Where C referenced to vortex center which represents the food source in shoveler ducks example. vortex behavior assumed that the previous three patterns combined to make the velocity vector

$$\vec{V}(t + 1) = S_i + A_i + C_i + w\vec{V}(t) \quad (13)$$

Where $w\vec{V}(t)$ represents propulsion force of the individual and w is the propulsion force factor which decaying with time and it's value between $[0 \ 1]$. Every individual needs to perform the next step by adding his step(velocity) to his position.

$$\vec{X}_i(t + 1) = \vec{X}_i(t) + \vec{v}_i(t + 1) \quad (14)$$

All of the previous equations make the individual move circular(vortex) motion but vortex cannot cover all directions so we have to make some small disturbance by mutate one dimension of some small random individuals to change vortex direction .

2. swimming: after performing cycling phase and all individuals are around specific point they begin to take the same direction of the center of vortex which determine by the difference between the location of the individual that has a fitness better than the center fitness (Gbest) and the center location (C).

$$\vec{c}_i(t + 1) = Gbest - C \quad (15)$$

$$\vec{X}_i(t + 1) = \vec{X}_i(t) + \vec{C}_i(t + 1) \quad (16)$$

Where $\vec{c}_i(t + 1)$ is center velocity vector, $\vec{X}_i(t + 1)$ is the position of ith individual including center.

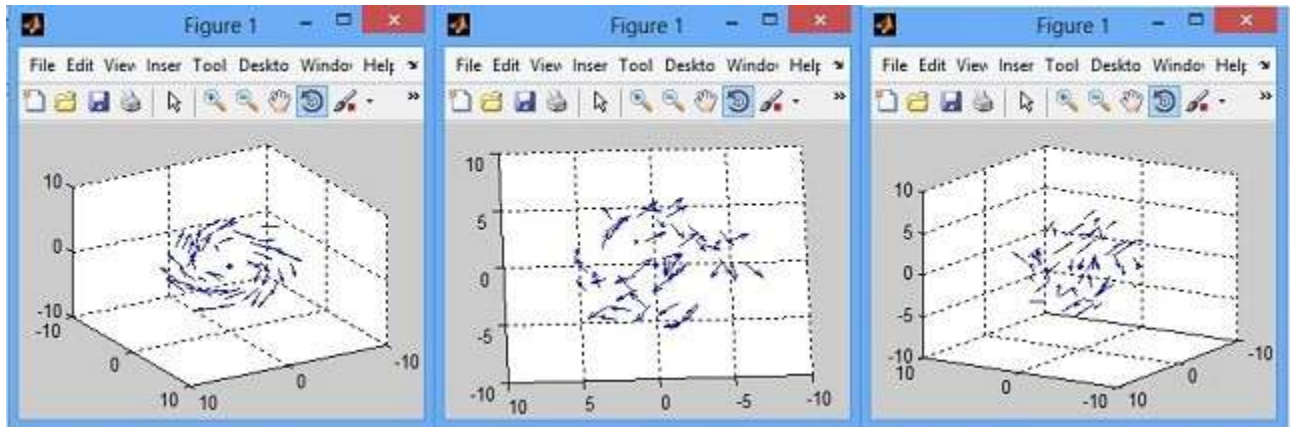


Fig. 10. Circling phase (vortex shape) 3D simulation from multiple perspectives. Interaction range=10 number of individuals=50 $w=1$ note that align and separation with neighbors are

normalized and weighted by distance, individuals have maximum speed equal 1 in every dimension

3. Foraging and feeding behaviors:

Any algorithm needs two strategies (exploration and exploitation) here in this model we will use the (p.vortex) technique for exploration where individuals increase their propulsion force to skip local minimum and reduce it in food location so propulsion force calculated as follows:

$$w = \begin{cases} M & iter < FI \\ w - \left(\frac{M-MI}{MT-FI} \right) & iter \geq FI \end{cases} \quad (17)$$

where iter is the iteration number, FI is the iteration that algorithm will perform exploitation after it (Feeding iteration), M refers to maximum propulsion force, MI is minimum propulsion force, M, MI values are between [0 1] and MT is the maximum iteration. By this equation some of individuals will have high value of propulsion force while algorithm explore the search space and all individuals have low value from the reduce radius iteration and decreasing until its value be equal to minimum propulsion force.

$$r = \begin{cases} val & iter < FI \\ r * (1 - DRR) & iter \geq FI \end{cases} \quad (18)$$

Where r is the radius of individual sensory range, val is the value of r before feeding iteration, and DRR is decrease radius rate. after while from performing algorithm the vortex reach the best promising area so the algorithm begin to transform to exploitation by reducing vortex radius which has the same value of sensory radius of individuals and that make individuals get more closer to each other and enhance solution the reduce radius step calculated as follow:

Algorithm 1 vortex swarm optimization

```

1: procedure VORTEX SWARM OPTIMIZATION
2:   while iteration < MaxIteration do
3:     circling(Radius,Particles,W)
4:     Swiming(Particles)
5:     Radius ← eq(18)
6:     W ← eq(17)
7:   end while
8: end procedure

```

Algorithm 2 Perform circling phase

```

1: procedure CIRCLING
2:   for i ← 1 to ParticlesNmbers do
3:     neighbours ← GetNeighbours(i,Particles,Radius)
4:     A ← Alignment(i,neighbours)eq(11)
5:     S ← Seperation(i,neighbours)eq(10)
6:     C ← AttractToCenter(Radius)eq(12)
7:     Velocity[i] ← W *Velocity[i] + A + S +C
8:     Position[i] ← Position[i] +Velocity[i]
9:   end for
10:  mutateSomeParticles(Particles)
11:  Evaluate(Particles)
12: end procedure

```

Algorithm 3 Perform Swimming Phase

```

1: procedure SWIMMING
2:   while BestFitness < CenterFitness do
3:     CenterVelocity ← BestPosition –CenterPosition eq(15)
4:     CenterPosition ← CenterPosition +CenterVelocity
5:     for i ← 1 to ParticlesNumbers do
6:       Position[i] ← Position[i] +CenterVelocity eq(16)
7:     end for
8:     BestFitness,BestPosition ← EvaluateParticles(ParticlesPositions)
9:     CenterFitness ← EvaluateCenter(CenterPosition)
10:    if NumOfSteps > MaximumSwimSteps then
11: break
12:   else
13: NumOfSteps ← NumOfSteps + 1
14:   end if
15: end while
16: end procedure

```

7.2 benchmark functions

here we will test our new algorithm on some benchmark functions [TABLE I] benchmark testing functions are divided to two groups unimodal [f1, f2],multi-modal [f3, f4, f5] where unimodal have single optimum and multi-modal have more than one optimum but one is global optimum and the rest are local optimum, or verification the results compare with three well-known algorithms [20] :Dragonfly algorithm, particle swarm optimization, Genetic algorithm each algorithm run 30 times over all the functions with 10 Dimensions , 30 individuals perform

algorithm and 500 iteration for each run, average and standard deviation of best solutions are calculated.

7.3 experimental parameters

Sensory range (r)=1.e-4, max swim steps=500, Feeding Iteration=300, mutation probability=0.05, Maximum propulsion force=0.2, Minimum propulsion force=0.05, Decrease radius rate=0.9

Results and conclusion

in [Table2] the results of the vortex swarm optimization proposed and shows advantages in unimodal and multimodal functions . The proposed swarm intelligence algorithm showed these results due to the effective mechanism of the new algorithm. Vortex swarm optimization inspired from vortex collective behavior where vortex shape forms by individuals in the group. Exploitation and exploration are developed by two behaviors (foraging, feeding)to get a good results in multimodal and unimodal benchmark functions. By increase propulsion force ,cycling and changing vortex direction the algorithm not stuck in local minimum and by reduce interaction range the algorithm begin to exploit around center , the algorithm was tested over five benchmark functions and compared with another effective algorithms and show a good results.

Table 1. [Benchmark functions]

<i>Function number</i>	<i>Function</i>	<i>Dim</i>	<i>Range</i>
f ₁	$f_1(\mathbf{x}) = \sum_{i=1}^n x_i^2$	10	[-100,100]
f ₂	$f_2(\mathbf{x}) = \sum_{i=1}^n x_i + \prod_{i=1}^n x_i $	10	[-10,10]
f ₃	$f_3(\mathbf{x}) = 10n + \sum_{i=1}^n (x_i^2 - 10\cos(2\pi x_i))$	10	[-5.12,5.12]
f ₄	$f_4(\mathbf{x}) = 1 + \sum_{i=1}^n \frac{x_i^2}{4000} - \prod_{i=1}^n \cos(\frac{x_i}{\sqrt{i}})$	10	[-600,600]
f ₅	$f_5(\mathbf{x}) = -a.\exp(-b\sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2}) - \exp(\frac{1}{n}\sum_{i=1}^n \cos(cx_i)) + a + \exp(1)$	10	[-32,32]

Table 2. [results]

Function number	SVO		DA		PSO		GA	
f1	avg	std	avg	std	avg	std	avg	std
F2	4.93E-43	2.69515E-42	2.85E-18	7.16E-18	4.2E-18	1.31E-17	748.5972	324.9262
F3	5.92119E-17	3.24317E-16	1.49E-05	3.76E-05	0.003154	0.009811	5.971358	1.533102
F4	0.132661208	0.344003018	16.01883	9.479113	10.44724	7.879807	25.51886	6.66936
F5	0.07639295	0.029420366	0.193354	0.073495	0.083463	0.035067	7.719959	3.62607
F6	7.49E-14	5.35562E-14	0.23103	0.487053	0.280137	0.601817	9.498785	1.271393

References

- Couzin, Iain D., et al. "Collective memory and spatial sorting in animal groups." *Journal of theoretical biology* 218.1 (2002): 1-11.
- Cziro'k, Andra's, and Tama's Vicsek. "Collective behavior of interacting self-propelled particles." *Physica A: Statistical Mechanics and its Applications* 281.1-4 (2000): 17-29.
- Canizo, J. A., J. A. Carrillo, and J. Rosado. "Collective behavior of animals: Swarming and com- plex patterns." *Arbor* 186 (2010):1035-1049.
- Sinkovits, Daniel. "Flocking behavior." University of Illinois (2006).
- T. Vicsek, "Novel Type of Phase Transition in a System of Self-Driven Particles," *Phys. Rev. Lett.* 75, 1226 (1995).
- Colorado Zuluaga, Gabriel Jaime. "Why animals come together, with the special case of mixed- species bird flocks." *Revista EIA* 19 (2013):49-66.
- Craig W. Reynolds. *Flocks, herds, and schools: A distributed behavioral model*. In *Computer Graphics*, pages 25–34, 1987.
- Craig W. Reynolds. *Boids, Background and Update* . <http://www.red3d.com/cwr/boids/>, 2007.
- [Online; accessed 14-April-2011].
- Vicsek, Tamas. "A question of scale." *Nature* 411.6836 (2001): 421.
- Vicsek, Tama's, et al. "Novel type of phase transition in a system of self-driven particles." *Physical review letters* 75.6 (1995):1226.
- Partridge, Brian L., and Tony J. Pitcher. "The sensory basis of fish schools: relative roles of lateral line and vision." *Journal of comparative physiology* 135.4 (1980): 315-325.
- Ben-Jacob, Eshel, Inon Cohen, and Herbert Levine. "Cooperative self-organization of microorganisms." *Advances in Physics* 49.4 (2000):395-554.
- Couzin, Iain D. "Collective cognition in animal groups." *Trends in cognitive sciences* 13.1 (2009): 36-43.
- Couzin, Iain D., and Jens Krause. "Self-organization and collective behavior in vertebrates." (2003): 1-75.
- Thomas, W. Beach. "Thought Transference (or what?) in Birds." *Saturday review of politics, literature, science and art* 152.3960 (1931): 362-362.

Couzin, Iain. "Collective minds." *Nature* 445.7129 (2007):715.

Delcourt, Johann, Nikolai WF Bode, and Mathieu Denoe'l. "Collective vortex behaviors: Diversity, proximate, and ultimate causes of circular animal group movements." *The Quarterly review of biology* 91.1 (2016):1-24.

Bode, Nikolai WF, and Johann Delcourt. "Individual-to-resource landscape interaction strength can explain different collective feeding behaviours." *PloS one* 8.10 (2013):e75879.

Cziro'k, Andra's, et al. "Formation of complex bacterial colonies via self-generated vortices." *Physical Review E* 54.2 (1996): 1791.

Mirjalili, Seyedali. "Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems." *Neural Computing and Applications* 27.4 (2016): 1053-1073.

Costanzo, A., and C. K. Hemelrijk. "Spontaneous emergence of milling (vortex state) in a Vicsek- like model." *Journal of Physics D: Applied Physics* 51.13 (2018): 134004.

Couzin, Iain D., and Nigel R. Franks. "Self-organized lane formation and optimized traffic flow in army ants." *Proceedings of the Royal Society of London B: Biological Sciences* 270.1511 (2003): 139-146.

Bazazi, Sepideh, et al. "Vortex formation and foraging in polyphenic spadefoot toad tadpoles." *Behavioral ecology and sociobiology* 66.6 (2012): 879-889.

Hemelrijk, Charlotte K., and Hanno Hildenbrandt. "Schools of fish and flocks of birds: their shape and internal structure by self-organization." *Interface focus* (2012): rsfs20120

Design Obfuscation for IP Security Applications

Ahmad M.H. Albatsha

Abstract:

The aim of this study is to give a brief overview of contemporary approaches (i.e., design/hardware obfuscation techniques) to protecting design (hardware) from various security threats – reverse engineering, IP piracy, tampering. The paper describes several existing and acknowledged technical solutions for design obfuscation as based on the analysis of patents published in the recent years. Eventually it substantiates the argument that it is essential to explore new hardware-efficient methods to combat the existing hardware attacks.

Keywords:

Design obfuscation; Integral Circuit (IC); Intellectual Property (IP) protection; reverse engineering; IP piracy; hardware Trojan

1. Introduction:

Recently, globalization and technological development have led to changes in attitudes toward hardware, which, apart from software, is now vulnerable to security threats. Integrated circuits (ICs) contain intellectual property (IP), which is critical for the stable functioning of financial, national defense and civil institutions. Illegal gaining and use of confidential data pose a threat to national security and economy. Currently, a growing number of commercial and government organizations are making use of different means of protecting the confidentiality of ICs. Among these means are design (hardware) obfuscation techniques.

Obfuscation is a method that transforms into a design, which is functionally equivalent to the original one, but at the same time considerably complicates its comprehension and reverse engineering. Design (hardware) obfuscation is related to electronic circuits (combinational, sequential, system-on-chip) and is concerned with the protection of semiconductor intellectual property. In the context of design (hardware), IP relates to “a reusable unit of logic, cell, or chip layout design that is either licensed to another party or owned or used solely by a single party.” (Shakya, Tehranipoor, Bhunia, Forte 2017: 6).

Design (hardware) obfuscation varies in technique and scope, depending on the threat model and the level of abstraction. Methods of design obfuscation are generally regarded at three levels: at the register-transfer level, at the layout level, and at the gate/netlist level (Becker, Fyrbiak, Kison 2017; Forte, Bhunia, Tehranipoor 2017).

* National Research Nuclear University MEPhI Moscow, Russia

It is impossible to find another universal criterion to incorporate all methods that are so diverse in operation and meaning. Yet it is essential to take note of emerging techniques, as there are still various limitations for different existing approaches, especially with regard to security considerations (Becker, Fyrbiak, Kison 2017).

2. Problem definition:

Due to the global distribution of the IC (Integrated Circuit) design flow, it is hard and sometimes even impossible to maintain trust between different organizations (or entities) in the production chain, which can be located in different parts of the world. The semiconductor supply chain consists of the IP (Intellectual Property) vendor, design house, DFT (design-for-test), foundry, assembly and market (for a detailed description see Shakya, Tehranipoor, Bhunia, Forte 2017: 7-9). Consequently, at any of these stages, hardware can become prone to such attacks as reverse engineering (i.e., recovering the IP by the malicious party), tampering (i.e., introducing vulnerabilities or backdoors into the design) and IP piracy (i.e., illegal or unlicensed use of IPs). Because of these attacks, the semiconductor industry loses up to \$4 billion each year (Sinanoglu, Pino, Rajendran, Karri 2014), thus making IC security one of the major issues in IP protection.

3. Objectives of the paper:

The aim of this paper is to give a brief overview of contemporary approaches (i.e., design/hardware obfuscation techniques) to protecting ICs from various security threats – reverse engineering, IP piracy, tampering. It describes several existing technical solutions for design obfuscation as based on the analysis of patents published in the last 7 years (from 2011 to 2018).

1. “Protection of intellectual property (IP) cores through a design flow” (2011, inventors: R.S. Chakraborty, S. Narasimhan, S. Bhunia);
2. “System and methods for silencing hardware backdoors” (2012, inventors: L. Sethumadhavan, A. Walksman);
3. “Systems, processes and computer-accessible medium for providing logic encryption utilizing fault analysis” (2014, inventors: O. Sinanoglu, Y. Pino, J. Rajendran, R. Karri);
4. “Methods and systems for preventing hardware Trojan insertion” (2015, inventors: M. Tehranipoor, K. Xiao);
5. “Post design integrated circuit manufacturing obfuscation” (2018, inventors: N. Levine, A. Gahoonia, J. Lloyd, D.W. Pentrack).

* National Research Nuclear University MEPhI Moscow, Russia

The paper explores recent patents, which describe new technical solutions to security applications in terms of their purpose, basic features and efficacy (as claimed by the inventors).

4. Data and analysis:

Among methods that can resist reverse engineering and IP piracy attacks is logic obfuscation. Logic encryption of hardware (or logic obfuscation) means concealing the functionality of the hardware. Logic obfuscation hides the functionality and the implementation of a design by inserting some additional circuit element into the original design: either additional logic states (in a sequential logic encryption), or XOR/XNOR gates (in a combinational logic encryption), or memory elements (Sinanoglu, Pino, Rajendran, Karri 2014). The inventors of the patent claim that logic encryption techniques can meet a security criterion, however this approach can be susceptible to several attacks including the Brute force attack (when the attacker applies all combinations of key bits and tries to figure out the correct key), the Collusion attack (when the attacker uses the key meant for IC_X on IC_Y), and the LEC removal attack (when the attacker removes the LECs from the encrypted netlist and randomly replaces them with a buffer or an inverter).

A countermeasure for sequential circuit protection is state obfuscation, the main purpose of which is to prevent the adversary from obtaining the clear state transition graph. The method HARPOON (from HARdware Protection through Obfuscation Of Netlist) is aimed at IP protection through obfuscation of netlist (Chakraborty, Bhunia 2009). The main principle is to add a key-controlled obfuscation mode before the finite state machine (FSM) enters the normal mode. The technique is capable of providing high levels of design obfuscation at nominal area overhead under delay constraint. However, the hardware obfuscation method HARPOON is shown elsewhere to be vulnerable to REFSM and fault injection attack (Meade, Zhao, Zhang, Pan, Jin 2017).

The patent invented by R.S. Chakraborty, S. Narasimhan, S. Bhunia (2011) suggests a method for protecting an IC chip design. The suggested method includes the following technical steps: a circuit description of an IC core, which comprises a set of nodes, is stored in the memory, and a plurality of modification nodes is selected from the set of nodes. A sequential structure inserted into the circuit description provides a modified circuit description, and the sequential structure utilizes the plurality of modification nodes as inputs. The modified circuit description is afterwards stored in the memory. The inventors argue that the obfuscation technique described in the patent assures that the components have gone through a secure and piracy-proof design flow, which has various benefits at each level of design and manufacture.

* National Research Nuclear University MEPhI Moscow, Russia

The next two technical solutions address methods for preventing tampering, the former being focused on the activation of hardware backdoors installed in a digital circuit (Sethumadhavan, Walksman 2012). The idea behind this solution is to make the backdoor design problem “intractable” to the attacker: inputs that are supplied to the hardware units at runtime are scrambled, thus making it impossible for malicious components to acquire the necessary data. The authors suggest three techniques, which in concert disable backdoor triggers: power resets (preventing untrusted logic from knowing that a large amount of time has passed since start-up), data encryption (XOR encryption or addition by a random value) and data reordering/insertion (benign reordering of sequences of inputs so as to preserve correctness but to prevent sequences from being deterministically supplied by an adversary). These methods are able to protect from such attacks as timebombs, single-shot cheat codes and sequence cheat codes. They can be applied to a wide variety of hardware components, e.g. it can be integrated into contemporary microprocessor designs.

Another technical solution aimed at protecting the design from hardware Trojan insertion was presented by Tehranipoor and Xiao in 2015 (Tehranipoor, Xiao 2015). The technique consists in using a plurality of built-in self-authentication (BISA) filler cells to fill unused spaces in an IC layout design. An example method can comprise determining unused space in an IC, selecting a plurality of BISA filler cells based on the determined unused space, and placing the selected plurality of BISA filler cells onto the unused space. BISA filler cells are shown to be immune to different attacks that attempt to create space for Trojan gates via removal, redesign, resizing or bypassing. The suggested methods and systems can be operational with numerous other general purpose or special purpose computing system environments or configurations (e.g., personal computers, server computers, laptop devices, multiprocessor systems, programmable consumer electronics, minicomputers etc.); they can employ Artificial Intelligence techniques such as machine learning and iterative learning.

The most recent patent (Levine, Gahoonia, Lloyd, Pentrack 2018) targets IP protection in the chain management of ICs. According to the suggested method, IC design layout files are partitioned into one or more design layout files, which contain data for front-end-of-line (FEOL) sub-circuits, and one or more design layout files, which contain data for back-end-of-line (BEOL) sub-circuits. A completed traditional IC design file is partitioned (morphed) into multiple design files, so that one design file will contain all of the layout information for a FEOL sub-circuit and multiple single layout layer design files will be created for each of the required BEOL layout layers. The FEOL sub-circuit design file and each of the BEOL single layer sub-circuit design files are sent to separate foundries to be manufactured on individual substrates.

* National Research Nuclear University MEPhI Moscow, Russia

In this way, no individual file alone contains sufficient information to deduce the overall function of the particular IC. As a result of the application of the method, the inventors claim to provide significant integrity and confidentiality protection, for the technique scheme reduces risk of IP theft during the manufacture of ICs. The efficacy of this IP protection method lies in the unaccountably large number of combinations and permutations of un-interconnected logic gates, which would be faced by an adversary, who might intercept either the FEOL sub-circuit or the BEOL sub-circuit.

5. Conclusion:

Design obfuscation methods are in demand in industry in order to prevent IP piracy, reverse engineering and tampering. Some of the methods suggested by researchers (inventors) seem to solve their primary tasks – to combat malicious attacks. However, in most cases, they do not appear to be immune to all possible threats. There is no universal solution for all design obfuscation problems, and many questions still remain unanswered. Unfortunately, for the most part generally accessible information on the latest methods of reverse engineering is limited, for companies majoring in reverse engineering do not publish their methods (Becker, Fyrbiak, Kison 2017). Attackers may obtain extra knowledge and tools to perform reverse engineering, tampering and IP piracy attacks, that is why in many cases the true advantages of design obfuscation for IP protection are hard to estimate in practice. One of possible strategies could be to explore the levels of obfuscation and protection achieved by using various techniques in order to understand what capabilities and instruments reverse engineers have at their disposal. Therefore, it is essential to explore new hardware-efficient methods to combat the existing hardware attacks.

6. Future work:

Having analyzed the recent technical solutions and the inventors' claims of their efficacy, we will examine these techniques in terms of their implementation and evaluation based on such metrics as fault impact, k-security, Hamming distance, verification failure, entropy. The choice of metrics will depend on the method under analysis.

* National Research Nuclear University MEPhI Moscow, Russia

References:

- Becker G.T., Fyrbiak M., Kison C. (2017). Hardware obfuscation: techniques and open challenges. In: Bossuet L., Torres L. (eds) Foundations of Hardware IP Protection, pp. 105-123. Cham: Springer.
- Chakraborty R.S., Bhunia S. (2009). HARPOON: An obfuscation-based SoC design methodology for hardware protection. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. Vol. 28, No. 10, pp. 1493-1502.
- Chakraborty R.S., Narasimhan S., Bhunia S. (2011). Protection of intellectual property (IP) cores through a design flow. US Patent 2011/0113392 A1. Appl. No.: 12/942,675. Filed: Nov. 9, 2010. Pub. Date: May 12, 2011. 35 p.
- Forte D., Bhunia S., Tehranipoor M.M. (eds) (2017). Hardware Protection through Obfuscation. Cham: Springer. xii, 349 p.
- Levine N., Gahoonia A., Lloyd J., Pentrack D.W. (2018). Post design integrated circuit manufacturing obfuscation. US Patent 9940419. Appl. No: 15/139,625. Filed: Apr 27, 2016. Pub. Date: Apr 10, 2018. URL: <https://patents.justia.com/patent/9940419>, accessed on 01.11.2018.
- Meade T., Zhao Z., Zhang S., Pan D.Z., Jin Y. (2017). Revisit sequential logic obfuscation: attacks and defenses. 2017 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-4.
- Sethumadhavan L., Walksman A. (2012). System and methods for silencing hardware backdoors. US Patent 2012/0124393 A1. Appl. No 13/273,016. Filed: Oct. 13, 2011. Pub. Date: May 17, 2012. 23 p.
- Shakya B., Tehranipoor M.M., Bhunia S., Forte D. (2017). Introduction to Hardware Obfuscation: Motivation, methods and evaluation. In D. Forte, S. Bhunia, M.M. Tehranipoor (eds) Hardware Protection through Obfuscation, pp. 3-31. Cham: Springer.
- Sinanoglu O., Pino Y., Rajendran J., Karri R. (2014). Systems, processes and computer-accessible medium for providing logic encryption utilizing fault analysis. US Patent 2014/0375353 A1. Appl. No.: 13/735,642. Filed: Jan. 7, 2013. Pub. Date: Dec. 25, 2014. 14 p.
- Tehranipoor M., Xiao K. (2015). Methods and systems for preventing hardware Trojan insertion. US Patent 9,218,506 B2. Appl. No: 14/204,656. Filed: Mar. 11, 2014. Pub. Date: Dec. 22, 2015. 19 p.

* National Research Nuclear University MEPhI Moscow, Russia

A Comprehensive E-government Adoption Literature Review

Abeer Mosaad Ghareeb¹ Nagy Ramadan Darwish² Hesham A. Hefney³

Abstract-The potential benefits from effective adoption of e-government are really large. The majority of countries has reached or is evolving towards higher levels of e-government development. There are some examples of clever use of ICTs and e-government in developing countries. However, the effect and the spread of e-government in these countries has so far been less than expected. Developing countries have invested more in core government administration systems than in transaction with citizens, and the governmental institutions are still unable to deliver basic services with reasonable quality. Low usage rate of the available e-government services is one of the main obstacles to the success of e-government projects. This paper aims to provide a literature review in the area of e-government adoption to examine the state of the art, identify the essential factors that are likely to affect citizen adoption, understand citizen's usage behavior regarding e-government services and ultimately put the groundwork for the development of a citizen-centric e-government adoption model.

Keywords: electronic government; e-government; literature review; adoption models; adoption factors; citizen requirements.

1. INTRODUCTION

In recent years, electronic government (e-government) has captured the attention of researchers and practitioners. There are several journals and annual conferences devoted to e-government. E-government refers to government agencies' use of Information and Communication Technologies (ICTs) such as wide Area networks, the internet, and mobile computing that have the ability to transform relations with citizens, businesses, and other arms of government. These technologies can serve a variety of different ends: better delivery of government services to citizens; improved interactions with business and industry; citizen empowerment through access to information; or more efficient government management. The resulting benefits can be less corruption, increased transparency, greater convenience, revenue growth and/or cost reductions (United Nations, 2016).

E-government is seen as an evolutionary phenomenon. Its maturity can be expressed using sequential steps, for instance, from immature to mature with improved quality and advanced features such as one stop shop, personalization, online payment, and interoperability. E-government maturity model is a set of stages that determine the maturity of e-government applications. Many models have been proposed to assess e-services provided to citizens. The most famous models are Layne & Lee model (Layne & Lee, 2001), Andersen & Henriksen model (Andersen & Henriksen, 2006), and United Nations model (United Nations, 2014).

¹Computer and Information Sciences Department, Institute of Statistical Studies and Research, Cairo University, Egypt. abeer_mosad@yahoo.com

²Computer and Information Sciences Department, Institute of Statistical Studies and Research, Cairo University, Egypt. drnagyd@yahoo.com

³Computer and Information Sciences Department, Institute of Statistical Studies and Research, Cairo University, Egypt. hehefney@ieee.org

A comparison between 25 e-government maturity models has been done in (Fath-Allah, Cheikhi, & Al-Qutaish, 2014). The findings showed that models' stages numbers varies from two to six. Almost all of the models contain a stage related to the availability of the portal in the web. Different names used while they focus on the same aspects (Emerging information, Publishing, Online presence, and Cataloguing), a stage where citizens can interact with government (Interaction), a stage where the citizens can transact with governments (Transaction), and a stage that covers advanced features such as information sharing between agencies (Integration). Moving towards the higher stages with transactional and connected services can be difficult. Such moving requires robust data protection systems and secure data sharing across government agencies (United Nations, 2014).

The adoption term has been described as a simple decision of using or not using online services depending on some factors (Kumar, Mukerji, Butt, & Persaud, 2007). Similar definition to e-government adoption as the individual's decision on whether or not to use the technology based on perceptions of the technology (Gilbert, Balestrini, & Littleboy, 2004). According to (Kumar, Mukerji, Butt, & Persaud, 2007), the next stage of adoption would be how frequently services are actually used. One usage per year, which may be considered as adoption by some, would not translate into meaningful usage by governments or citizens. Scope of usage is another very important dimension of adoption. The other dimension includes preference of the online medium over other mediums of transactions with government.

The potential benefits from effective adoption of e-government, particularly in developing countries, are really large. E-government can contribute to development, make governments more transparent and accountable, improve disaster mitigation and management, help governmental institutions to enhance the quality of basic services in all sectors, and ultimately enhance citizen satisfaction with government performance. There are some examples of clever use of ICTs in developing countries such as mWater, which is a mobile and web platform for monitoring and regulating 252 water schemes in small towns such as in Senegal, Mali, Benin, and Niger (United Nations, 2018). However, effect and spread of e-government in developing countries has so far been less than expected. Developing countries have invested more in core government administration systems than in transacting with citizens and businesses (World Bank, 2016).

The aim of this paper is to review the prior studies and relevant publications about e-government adoption to get the important background information about the subject under study, to identify the suitable data-collection instruments, to avoid duplicating previous research, and to be aware of inconsistencies and gaps that may need extra research. The remainder of the paper is organized as follows. Section 2 provides our research methodology. Section 3 provides the literature review. Section 4 introduces the research findings. Finally section 5 gives the conclusion and future work.

2. RESEARCH METHODOLOGY

The literature review is a systematic, explicit, and reproducible method for identifying, evaluating, and interpreting an existing body of completed and recorded work produced by researchers, scholars, and practitioners (Fink, 2014). It provides a comprehensive overview and helps place that information into perspective (Green, Johnson, & Adams, 2001). International organizations' reports, Journals articles and conference proceedings were the data sources provided the input for our literature review. The diversity and quality of the publications involved in the present literature review is assumed to provide an adequate and dependable

Cairo University, Institute of Statistical Studies and Research

sample of existing research on e-government adoption. Preliminary reading of the papers' abstracts was conducted to locate potentially relevant studies. The selected studies were read and analyzed in depth to fill the predefined extraction forms which have been designed to summarize the relevant information from the reviewed empirical studies. We have excluded the studies not directly relating to e-government adoption and acceptance, and the studies with no data reported about them. In addition to, the studies with unclear methodology, unclear relationships between the dependent and independent variables, or the factors are divided into irrational dimensions.

3. LITERATURE REVIEW

In this section a literature review in the area of e-government adoption is provided in terms of technology adoption theories, the United Nations' surveys, and previously empirical studies. A great amount of work has been done but there is still much to offer to improve the situation.

3.1. Technology Adoption Models/Theories

Technology Acceptance Model (TAM) (Davis, 1989), Diffusion of Innovation Model (DOI) (Rogers, 1995) and the Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh, Morris, Davis, & Davis, 2003) are the well-known models developed throughout the years to study and investigate the effect of factors on the acceptance and use of technologies. Table1 summarizes the definition of the basic constructs of TAM, DOI, and UTAUT. Table 2 summarizes some limitations of these models. The importance of these technology adoption models is not denied. They provide useful insights and implications for understanding an individual's intention of using e-government information and services (Komba, 2013). Many researchers have introduced their own citizen adoption models which are constructed based on TAM, DOI, and/or UTAUT (Alsaif, 2013) (Alshehri, Drew, Alhussain, & Alghamdi, 2012) (Hu, Brown, Thong, Chan, & Tam, 2009) (Alomari, 2014) (Komba, 2013) and amended the inadequacies of their original forms by including additional constructs such as trust and web site design.

Table 1 Basic Constructs of TAM, DOI, and UTAUT

Model	Basic Constructs	Definition of Basic Constructs
TAM	Perceived usefulness	The degree that users believe that a particular system facilitate their activity.
	Perceived ease of use	The degree that users believe that using the system is easy.
	Relative advantage	The degree to which an innovation is seen as being superior to its predecessor.
DOI	Complexity	The degree to which an innovation is perceived as relatively difficult to understand and use.
	Compatibility	the degree to which an innovation is seen as to be compatible with existing values, beliefs, experiences, and needs of adopters.
	Trialability	Refers to how easily a new technology can be used experimentally.
UTAUT	Observability	Refers to the extent to which the innovation provides visible and tangible results.
	Performance expectancy	The degree to which an individual believes that using the system will help him or her to attain gains in job performance.
	Effort expectancy	refers to the degree of ease associated with use of the system
	Social influence	the degree to which an individual perceives that important others believe he or she should use the new system
	Facilitating conditions	the degree to which an individual believes that an organizational and technical infrastructure exists to support use of the system

Table 2 Limitations of TAM, DOI, and UTAUT

Model	Limitations
TAM	Lack of many important factors that have direct impact on behavior in term of intention to use (Alghamdi & Beloff, 2014). Represents subjective user assessments of a system.
DOI	Certain key constructs in DOI are analogous to the constructs in TAM. Relative importance is similar to perceived usefulness and complexity is similar to perceived ease of use (Colesca & Dobrica, 2008) (Komba, 2013). The observability construct can be implicitly integrated into the relative advantage construct (Alghamdi & Beloff, 2014).
UTAUT	Does not address some very important constructs relating to e-government context such as trust and awareness. Grouping and labeling of items in UTAUT is problematic (Alghamdi & Beloff, 2014).

3.2. United Nations' E-government Surveys

Since 2001, the United Nations has issued the United Nations e-government survey each two years to assess the e-government development status of the 193 United Nations member states using an e-government development index (EGDI). EGDI is a composite measure of three important dimensions of e-government, namely: provision of online service; telecommunication connectivity; and human capacity. According to the 2018 Survey, Denmark ranks highest among the top 10 countries leading e-government development, followed by Australia and the Republic of Korea. The remaining countries in the top 10 are the United Kingdom, Sweden, Finland, Singapore, New Zealand, France and Japan. The highest regional EGDI is Europe (0.7727), followed by Americas (0.5898), Asia (0.5779), Oceania (0.4611), and finally Africa (0.3423). Investigating previous trends, there has been no change in regional positions since 2003 (United Nations, 2018).

By 2014, all 193 member states of the United Nations had national website (United Nations, 2014). The average world EGDI has been increasing from 0.47 in 2014 to 0.55 in 2018 due to continuous enhancement of online services provision. This implies that, globally, there has been steady progress in enhancing e-government and public service provision online. Countries continue to advance to higher levels of e-government. 18 countries across the regions improving from Low-EGDI level to Medium-EGDI level. This is followed by 17 countries that progressed from Medium-EGDI level to High-EGDI level and 11 moving from High-EGDI level to Very-High-EGDI level. Ghana is the only African country that moved from Medium-EGDI level to High-EGDI level. Delivery of services through mobile applications is growing fastest in the education, employment and environment sectors (United Nations, 2018).

Even with this progress, e-government divide, similar to digital divide, continues. The majority of the world' population remains offline. Almost one-half or 48 per cent of the world's population is estimated to have used the internet in 2017 and what people do with their access is a key difference in whether users take full advantage of e-government services (United Nations, 2018). Adoption and use of e-government services is still rather limited in most nations (United Nations, 2014). Although the number of countries providing income tax services online increased from 73 countries in 2014 to 114 in 2016 low usage by end users still one of the main obstacles to the advance of the e-government projects (United Nations, 2016). The improvement of business environment has remained a priority for the period 2014-2016, with 37 countries introducing the online business registration, the application for the number of registration and licenses still remain at low levels (United Nations, 2016). These outcomes tend to show that there is a substantial general lack of understanding about e-government services usage. The governments need to avoid the technocratic approach that result in online services with minimal

usage (United Nations, 2016). Rather they need to focus more on finding ways to enhance usage rates of the available e-government services to achieve the full potential of e-government.

3.3. Review of Empirical Studies

Several empirical studies have been accomplished on e-government adoption topic. The review of literature identified a number of the factors influencing the adoption of e-government information and services, especially, from the perspective of citizen. The discussion below presents the observed findings of some of the e-government adoption studies.

Investigation of the effect of the Website Quality on the acceptance of using e-government services in the Saudi Arabia has been done in (Alshehri, Drew, Alhussain, & Alghamdi, 2012). The findings have revealed that Website quality is one of the main significant factors. Effort expectancy, performance expectancy, and facilitating conditions directly affect the usage behavior of e-government services. In (Alateyah, Crowder, & Wills, 2013), an integrated model, based on DOI with some added variables, has been developed to examine adoption of e-government systems in Saudi Arabia by its citizens. The results of most of the identified factors have been seen adequate. The study also found that there is a problem with delivering e-government services due to the postal infrastructure. The Saudi citizens do not have a clear address, such as house number or a clear street name, or secure postal address which usually available in front of their house or a box beside their doors. Although there are post office Boxes available for the citizen rent, not many Saudis have one, which makes the communication between the citizens and the government difficult. Therefore, the mail service system in Saudi Arabia needs to be developed and improved to become a certified contact.

The findings, of an empirical examination of the awareness and adoption of the ‘Government Gateway’ amongst UK citizens, revealed that citizens with home broadband access are more likely to be aware of and adopt e-government services. Demographic characteristics of citizens such as the age, gender, education, and social class have an imperative role in explaining the citizen’s awareness and adoption of e-government services in the household. As increases in the citizens’ age ranges occurred, there was more awareness of and adoption of the Government gateway. An interesting obtained result was that the adoption rate declines considerably after 54 years. Males are more bound to drive adoption. Awareness and adoption of e-government services decreases as the social class of respondents reduces. And the higher the income of respondents the more probable it was that the e-government services would be adopted (Choudrie & Dwivedi, 2005).

In (Hu, Brown, Thong, Chan, & Tam, 2009), the main drivers of service quality and subsequent continuance intention of e-Tax, has been analyzed. The results showed that security, convenience, and perceived usefulness are the key determinants of e-Tax service quality. Perceived ease of use was not a significant predictor, which can likely be attributed to the context of the study—i.e., Hong Kong is one of the most technologically sophisticated societies and further, the respondents in the survey were fairly young, a group to whom perceived ease of use tends to be less important. The study in (Srivastava & Teo, 2005) revealed that Singapore government is systematically addressing both citizen trust on government and citizen trust on technology and is slowly but steadily moving towards a regimen of high trust on both dimensions, which offer a set of lessons learned for other nations to emulate.

The authors in (Alomari, 2014) integrated factors from TAM and DOI. In addition to some additional factors and indicated that trust in the internet, website design, religious beliefs, internet and computer skill confidence, word of mouth, resistance to change, perceived usefulness, relative advantage and complexity are the main factors that should be considered when addressing the topic of e-government services' adoption in Jordan. Furthermore, twelve environmental factors, that Jordanian citizens perceive to have an influence on the adoption expectancy of e-government services, have been explored in (Abu-Shannab, 2012). The twelve factors have been organized around three dimensions: the social dimension, the infrastructure dimension and the governmental dimension. The social dimension was the highest in Jordanian feeling of the obstacles against the success of e-government, which mainly trust and privacy. In (Meftahi, Gharleghi, & Samadi, 2015), the results of a regression test showed strong evidence of a significant relationship between culture, awareness and trust and adoption of e-government in Bahrain. The effectiveness of marketing e-government services in Egypt, by examining the views of students at both private and public universities, has been analyzed in (El-Baradei, Shamma, & Saada, 2012). The study concluded that the government was to a large extent not successful in marketing the adoption of e government services.

The findings of an examination of the channel choice and digital divide in Egypt revealed that contacting the government through traditional channels is the most frequently used methods. In-person visit was the most used channel, with 92.9% of the survey sample using it and the phone is the second most method, with 17.2% of the citizens using it (Reddick, AbdelSalam, & Elkadi, 2012). Analyzing the case of Romania has shown that the citizen's higher perception of usefulness and ease of use, quality and trust on e-government services directly improve the citizen' satisfaction and implicitly the level of e-government adoption. The study also aimed to determine the influence of demographic variables on perceived usefulness and ease of use and ultimately on the adoption process. It has been found that educational level and age are the most significant variables (Colesca & Dobrica, 2008). The study in (Rokhman, 2011) aimed to find out how the acceptance of Indonesian Internet users to e-government services. Relative advantage and compatibility were confirmed as useful factors to predict intention of use of e-government. Otherwise the variables of image and ease to use are not confirmed.

An assessment of the current situation of e-government adoption in Tanzania and identification of the critical factors affecting use of e-government services have been done in (Komba, 2013) in order to provide an e-government adoption model within which new initiatives might be evaluated. The proposed model integrated factors from TAM, DOI, demographic characteristics, information needs, and a set of external factors. The findings shown that, the current state of e-government in the selected districts of Tanzania is characterized by web presence stage. Numerous enhancing factors and challenges for e-government were noted. In (Lin, Fofanah, & Liang, 2011), the results of testing the validity of TAM in the e-government setting in Gambia revealed that the basic constructs of the TAM have strong influences on user-intention towards e-government products. It is observed that perceived ease of use and information quality have strong impact on perceived usefulness toward using e-government.

The researchers in (Hung, Tang, Chang, & Ke, 2009) investigated the effect of a set of factors on the employee's intention to use Electronic Document Management System (IDMS) in Taiwan. IDMS is the most popular intergovernmental service in e-government projects. The study showed that the increasing effective of training over trust and perceived ease of use is more important than perceived usefulness. In (Al-Mamari, Corbitt, & Gekara, 2013) the major

motivating factors for the Omani government to implement e-government projects have been highlighted. Economic factor played a stronger motivating role followed by institutional necessities to comply with international standards. Motives of service quality and IQ were regarded as complementary to the economically-forced strategy of e-government implementation within a national move towards sustainable knowledge-based economy.

A data extraction form has been designed to summarize the relevant information from the reviewed empirical studies. The extracted data are organized into two tables. Table 3, contains the context of the study, the base models/theories, dependent and independent variables, and the findings. Table 4 contains the description of the methodologies in terms of research strategy, instrument, sampling, validity and reliability and analysis technique. The vast majority of the reviewed empirical studies employed the quantitative approach. The questionnaire and interview were the main data collection instruments. Except (Alateyah, Crowder, & Wills, 2013) (Abdelkader, 2015) and (Al-Mamari, Corbitt, & Gekara, 2013), the sample size ranged from 140 to 1191. The most used analysis technique was the multiple regression analysis. Other employed analysis techniques were the structural equation modeling, sample t-test and Hermeneutics.

4. RESEARCH FINDINGS

Some important insights and messages can be derived from this literature review that offer some guidance and implications.

- Little research has been done on e-government adoption from government perspective. The most critical issues that should be considered when discuss e-government adoption from government perspective are funding and motives of policy makers to adopt e-government initiative.
- The resistance or rejection of e-government systems within government institutions may negatively affect spread of the phenomenon of e-government. The most important factors, which help stopping obstacles the adoption e-government services from inside the government agencies, have not been adequately discussed.
- The majority of the reviewed studies investigated the effect of a number of adoption factors, as independent variables, on the intention to use e-government services, as a dependent variable. However, intention to use only does not describe the full spirit of adoption.
- The lack of interest among citizens to use e-government services either because they don't aware with existence of these services or they know their existence but there are obstacles that hinder the usage.
- There is a discrepancy in citizen usage rates of e-government services from one country to another with low usage rates observed in many developing countries such as Egypt (Abbassy & Mesbah, 2016), Jordan (Abu-Shannab, 2012), Oman (Al-Mamari, Corbitt, & Gekara, 2013), Bahrain (Meftahi, Gharlegghi, & Samadi, 2015) and Zambia (Bwalya, 2009). Even in developed countries such as Australia, Canada, and New Zealand, citizens use e-government services mostly to get information and not for transacting with government (World Bank, 2016).

- Some important adoption factors are involved in the reviewed studies such as perceived ease of use, perceived usefulness, trust, awareness, service quality, web site design, availability of resources, demographic characteristics, resistance to change, compatibility, and word of mouth. However, a number of significant factors that are expected to have influence on the adoption of e-government were missed in the literature. Examples of these missing factors are citizen satisfaction, delivery channels, payment methods, cost, and mandatory.
- Citizen satisfaction factor has been approximately absent from e-government adoption models proposed in the literature. Provision of accessible, transparent, and convenient public services will facilitate a good life to all population and enhance citizen satisfaction with public services. The governments must monitor citizens' satisfaction with existing services.
- The governments around the world have no choice not to adopt e-government. Why the government leaves it a choice for government employees and citizens. The ordinary citizens prefer non-digital service channels, such as traditional in-person visits to governmental offices, when both digital and non-digital service delivery channels are available. Being e-government services mandatory will force citizens to try them and detect their numerous benefits. The study in (El-Baradei, Shamma, & Saada, 2012) is the only one that mentioned the mandate factor and dealt with it as a recommendation, not as an explicit factor.
- Efforts are needed to promote the citizens to use e-government services by making ICTs accessible and affordable, offering user-friendly interfaces, considering the quality of the presented information, decreasing the risks of security breaches, localizing e-government applications, providing cheaper services, reducing service delivery time, employing convenient payment mechanisms, raising the awareness of the value of digital use, and forcing digital service delivery channels.
- Lack of e-government services quality measurement to assess the performance of the services delivered to citizens.
- Adoption process of e-government services is multi-criteria in nature. It involves a crowd of interrelated factors with varying degrees of influence. Multi-criteria decision making (MCDM) used widely for evaluation and ranking problems containing multiple criteria. However, the application of MCDM in the field of e-government adoption is still rare.

Table 3 Reviewed empirical studies

Study	Context	Base Models	Research Topic	Dependent Variables	Independent Variables	Findings
(Alsaif, 2013)	Kingdom of Saudi Arabia	UTAUT	Investigating the problem of low adoption of e-government from the citizen's perspective by determining the psychological, technical and social factors, moderated by socio-cultural value	Intention to use Usage behavior	Intentions to use Computer-self Efficacy Availability of Resources Information Quality System quality Awareness of the System Performance Expectancy Effort expectancy Compatibility Social influence Trust in Internet Trust in Government	Performance expectancy is the strong predictor of the intention to use e-government followed by the trust in the internet. Intention to use , computer self-efficacy and availability of resources found to be significant predictors of usage behavior
(Alshehri, Drew, Alhussain, & Alghamdi, 2012)	Kingdom of Saudi Arabia	UTAUT	Studying the effect of website quality on the actual usage of e-government services from the perspective of Saudi Arabian citizens	Use Behavior of e-government services	Performance Expectancy Effort expectancy Facilitating Conditions Website Quality Social Influence	Website quality is a main significant factor of e-government services adoption Effort expectancy, performance expectancy, and facilitating conditions contribute significantly to citizen adoption of e-government services. Social influence was insignificant.
(Alateyah, Crowder, & Wills, 2013)	Kingdom of Saudi Arabia	DOI	Factors influencing citizens' intention to adopt e government services in Saudi Arabia	Intention to adopt	Infrastructure computer literacy Demographic char. security service quality speed of delivery reliability DOI factors usability accessibility multi-lingual, awareness, culture	The result of most of the identified factors have been seen adequate The result of data analysis of the factor Multi-Lingual has been accepted by the result of the data analysis for the questionnaires applied to citizen and government's employee as well as from the background, even though the experts' result is seen not adequate.
(Alomari, 2014)	Jordan	DOI + TAM	Identifying the main factors related to e-government adoption by the public in Jordan	e-government adoption	Trust in internet Website design Religious beliefs Internet and computer skill Confidence Word of mouth Resistance to change Wasta (favoritism) Perceived usefulness Relative advantage	The main factors related to e-government adoption are trust in internet, religious beliefs, website design, internet and computer skill confidence, word of mouth, resistance to change, perceived usefulness, relative advantage, and complexity. Wasta was the only factor was dropped down from further analysis since it has recorded a weak loading.
(Abu-Shannab,	Jordan	-	Exploring the factors that Jordanian citizens perceive to	adoption expectancy	Twelve environmental factors organized around three	The social dimension is the highest in Jordanian feeling of obstacles against

Study	Context	Base Models	Research Topic	Dependent Variables	Independent Variables	Findings
2012)			have an influence on the adoption expectancy of e-government services		dimensions: The social dimension, the infrastructure dimension and the governmental.	success of e-government, which mainly trust and privacy. The lowest two factors are governmental ones, which mainly the resistance of public departments and adequacy of the legal framework
(Rokhman, 2011)	Indonesia	DOI	Finding out how the acceptance of Indonesian Internet users to e-government services	Intention to use	Relative advantages Image, Ease to use Compatibility	Two significant variables are relative advantage and compatibility. Two not-significant variables are image and ease to use.
(Abdul-Razak, Abu-Bakar, & Abdullah, 2017)	Malaysia		Examining the role of effort expectancy and social influence in explaining Malaysian government service acceptance of e-government services from the perspective of Malaysian government servants	Continuance intention to use	Ease of use Effort expectancy Social influence	The combination of effort expectancy and social influence were found to have significant impacts on continuance intention to use e-government
(Meftahi, Gharleghi, & Samadi, 2015)	Bahrain		The relationship among culture, awareness, and trust towards e-government adoption.	Intention to adopt	Cultural Awareness Trust	The three variables found to have a significant relationship with intention to adopt. Trust shows the highest impact, followed by awareness and then culture.
(Lin, Fofanah, & Liang, 2011)	Gambia	TAM	validity testing of TAM in e-government setting in Gambia and focuses on how Gambians behave differently than other e-government users	Gambia's e-Government usage intention	information quality perceived usefulness perceived ease of use attitude towards using behavioral intention	The core constructs of the TAMs have a significant and strong influence on Gambia's e-government usage intention.
(Choudrie & Dwivedi, 2005)	United Kingdom	-	Examining the citizens' awareness and adoption of e-government initiatives in the United Kingdom (UK), specifically the 'Government Gateway'	Awareness and adoption of the 'Government Gateway' amongst UK citizens	Demographic char.	Demographic characteristics of citizens such as age, gender, education, and social class have an imperative role in explaining citizen's awareness and adoption of e-government services in the household. Citizens with home broadband access are more likely to be aware of and adopt e-government services
(Colesca & Dobrica, 2008)	Romania	TAM	Determining the influence of demographic variables on perceived usefulness and ease of use and ultimately on the adoption process	Citizen satisfaction Adoption	Citizen satisfaction Perceived usefulness Perceived ease of use Perceived trust Perceived quality Demographic variables.	Educational level and age are the most significant demographic variables that influence on perceived usefulness and ease of use. Citizen's higher perception of usefulness, ease of use, quality and trust directly improve the citizen' satisfaction and implicitly the level of adoption.
(Hu, Brown, Thong, Chan, &	Hong Kong	TAM	Examining the determinants of service quality and continuance intention of e-Tax online services	Service quality Continuance intention	Perceived usefulness Perceived ease of use Security	Both service characteristics (i.e., security and convenience) and one of the technology characteristics (i.e.,

Study	Context	Base Models	Research Topic	Dependent Variables	Independent Variables	Findings
Tam, 2009)					Convenience of e-Tax Service quality	perceived usefulness, but not perceived ease of use) were the key determinants of service quality.
(El-Baradei, Shamma, & Saada, 2012)	Egypt	-	Examining the extent to which Egyptian government was successful in marketing its e government services to citizens			The government was to a large extent not successful in marketing the adoption of e-government services
(Reddick, AbdelSalam, & Elkadi, 2012)	Egypt		Examining channel use and digital divide in Egypt	Channel choice	Use internet daily from home Broadband internet access Dialup internet access Mobile phone internet access Age College educated Employed Female Urban Low economic status	There is a digital divide in the use of e-government by citizens. Major contact channel is the in-person visit to a government office. The second most commonly used channel is the phone.
(Abdelkader, 2015)	Egypt		Addressing two central e-government projects that are being undertaken in governmental organizations in Egypt and focusing on barriers that influence the implementation of e-services. These projects are Family Card System and the Ministry of Justice projects.			The study concluded that political will is a driver for all e-government implementation in Egypt, social considerations should be considered prior to technical aspects, a modular approach to rollout hug projects is best versus going for a big bang approach, and centralization ICTs applications save many costs.
(Srivastava & Teo, 2005)	Singapore	literature on trust	Examining the process by which Singapore government is promoting and maintaining trust for e-government among its citizens			Citizen trust on government is being generated by putting institutional trust building measures in place, taking feedback from citizens and having top leadership commitment and support for e-government. Citizen trust on technology is being engendered by a provision of relevant knowledge and skills and providing a comprehensive and effective legal system
(Mandari & Chong, 2018)	Tanzania	DOI	Determining the effect of innovation diffusion theory constructs and government support on rural farmers' behavioral intention to use m-government	Behavior intention	Relative advantage Ease of use Compatibility Image Government support	Ease of use was found to be the strongest determinant of behavior intention and this may be attributed to the fact that most of the rural farmers are less technology savvy. Age and gender were found to moderate the effect of relative advantage and ease of use.

Study	Context	Base Models	Research Topic	Dependent Variables	Independent Variables	Findings
(Komba, 2013)	Tanzania	TAM + DOI	Investigating the current situation and factors influencing access of e-government information and e-government adoption in Tanzania	e-government adoption	Information needs Demographic char. External factors TAM factors DOI factors	The current state of e-government in the selected districts of Tanzania is characterized by web presence stage. Enhancing factors include necessary knowledge and resources, awareness, confidence to use the website, availability and reliability of internet connections, possession of ICTs, up-to-date information in the website, education, income and social influence.
(Hung, Tang, Chang, & Ke, 2009)	Taiwan	TPB + TAM	Identifying the factors that determine employee acceptance of electronic document management system (EDMS) which is the most popular intergovernmental service in the e-government project (employee perspective)	intention to accept EDMS	perceived usefulness perceived ease of use Training, Trust personal innovativeness Compatibility External influence Interpersonal influence Self-efficacy Facilitating conditions	Perceived usefulness, perceived ease of use, training, compatibility, external influence, interpersonal influence, self-efficacy, and facilitating conditions are significant predictors of employees' intention to utilize EDMS. The study showed that the increasing effective of training over trust and perceived ease of use is more important than perceived usefulness.
(Al-Mamari, Corbitt, & Gekara, 2013)	Oman	Institutional theory, information quality, and quality of service models	Examining and highlight the key motivating factors for the Omani Government to implement e-government projects (government perspective)			The economic factor specific to Oman played a stronger motivating role second by institutional necessities to comply with international standards. Measures of service quality and IQ were regarded as complementary to the economically-forced strategy of e-government implementation within a national move towards sustainable knowledge-based economy.

Table 4 Employed methodologies in the reviewed empirical Studies

Study	Research Strategy	Instrument	Sampling	Validity & Reliability	Analysis Technique
(Alsaif, 2013)	Quantitative	Online questionnaire	692 Saudi Citizen	Factor Analysis Cronbach's Alpha	Multiple regression analysis
(Alshehri, Drew, Alhussain, & Alghamdi, 2012)	Quantitative	Questionnaire	400 Saudi Citizen randomly chosen from 3 big cities in different geographic regions: Riyadh, Jeddah and Abha.	Factor Analysis Cronbach's Alpha	Structural equation modeling using AMOS tools
(Alateyah, Crowder, & Wills, 2013)	Mixed	Interviews and questionnaires	31 Saudi Arabian government employees and 8 experts in the field of information technology	Cronbach's Alpha	One sample t-test. Accepted value has been identified as 3.5 and above.
(Alomari, 2014)	Quantitative	Two questionnaires were designed for the two government departments	356 Jordanian citizens	Factor Analysis Cronbach's Alpha	
(Abu-Shannab, 2012)	Quantitative	Questionnaire of 15 items	A sample of 259 teachers and employees	Factor Analysis Cronbach's Alpha	Multiple regression analysis
(Rokhman, 2011)	Quantitative	Online questionnaire	751 Indonesia Internet users	Factor Analysis Cronbach's Alpha	Logistic regression analysis
(Abdul-Razak, Abu-Bakar, & Abdullah, 2017)	Quantitative	Questionnaire	543 government servants at a selected public school and most of the sample comprised secondary school teachers.	Exploratory factor analysis Cronbach's Alpha	Multiple linear regression
(Meftahi, Gharlegghi, & Samadi, 2015)	Quantitative	Questionnaire	140 private sector employees	Cronbach's Alpha	Multiple regression analysis using SPSS
(Lin, Fofanah, & Liang, 2011)	Quantitative	Questionnaire	146 employees who were identified as having greater than average access to the internet or e-Government facilities. Through e-mail	Measures were constructed by adopting constructs validated by other researchers Cronbach's Alpha	Applying Structural equation modeling manually
(Choudrie & Dwivedi, 2005)	Quantitative	A postal self-administered questionnaire	358 households		Multiple regression analysis
(Colesca & Dobrica, 2008)	Quantitative	Online questionnaire	481 citizens	Cronbach's Alpha	Multiple regression analysis
(Hu, Brown, Thong, Chan, & Tam, 2009)	Quantitative	Online questionnaire	518 participants before and after they made use of the e-Tax service	Factor Analysis Cronbach's Alpha	Regression analysis and Post-hoc Analysis
(El-Baradei, Shamma, & Saada, 2012)	Mixed	Questionnaire	A sample of 363 students in two different types of university campuses; private universities represented by the American University in Cairo, and public universities represented by Cairo University and Alexandria University		Sample t-test Quantitative data were analyzed using SPSS. Qualitative data obtained through the open ended question were reviewed by the authors to discern main themes and ideas and complement the quantitative analysis

Study	Research Strategy	Instrument	Sampling	Validity & Reliability	Analysis Technique
(Reddick, AbdelSalam, & Elkadi, 2012)	Quantitative	Questionnaire	1191 adult citizens (18 years old and above) in Egyptian households having landline phones nationwide.		Descriptive statistics was used to determine the extent of citizens' use of each channel. This is followed by logistic regression and linear regression to examine the impact of digital divide variables on channel choice. Coding process using NVIVO
(Abdelkader, 2015)	Qualitative Case study	Semi-structured interviews	10 Semi-structured interviews with key senior officials involved in e-government services development in Egypt.		
(Srivastava & Teo, 2005)	Qualitative Case study	Case study			
(Mandari & Chong, 2018)	Quantitative	Questionnaire	407 Tanzanian rural farmers	Confirmatory analysis	factor Structural equation modeling
(Komba, 2013)	Triangulation approach	Questionnaire and semi-structured interviews	448 respondents for questionnaire survey + interviews with five policy makers	Factor Analysis	Multiple linear regression Quantitative data were analyzed statistically using SPSS software. Qualitative data were coded and analyzed using a computer database program called NUD.IST Vivo (Nvivo). Structural equation modeling
(Hung, Tang, Chang, & Ke, 2009)	Quantitative	Questionnaire	186 employees of real e-government's EDMS in Taiwan	Factor Analysis Cronbach's Alpha	
(Al-Mamari, Corbitt, & Gekara, 2013)	Qualitative case study	Semi-structured interviews	22 individuals who participated in setting the strategic plan of implementing e-government in Oman and others who participated in the actual implementation process from ten different government units		Hermeneutics technique using the Nvivoq software package

5. CONCLUSION AND FUTURE WORK

Based on the literature review and research implications, two important remarks can be concluded. Firstly, there is a need to propose e-government adoption models that illustrate the steps of adoption process from the point of view of citizen and contain the essential factors, including those that were lost in the previous work, which are expected to have a significant impact on the behavior of the citizen regarding e-government usage. Therefore, a citizen-centric e-government adoption model is currently developed by the current researchers. The potential model assumes that intention to use e-government services does not express the full spirit of adoption. The addition of actual use and potential use express the process and stages of adoption from the point of view of citizens. Secondly, evaluating the performance of the deployed e-government services, and discover their strong and weak points, is the upmost important. Some e-government services are more likely not to be used due to some reasons such as complications in usability or inconvenient payment method. Therefore, the current researchers are currently developing an approach that can be used to evaluate the performance of deployed e-government services through multi-criteria analysis.

REFERENCES

- Abbassy, M. M., & Mesbah, S. (2016). Effective E-government and Citizens Adoption in Egypt. *International Journal of Computer Applications*, 133(7), 7-13.
- Abdelkader, A. A. (2015). A manifest of barriers to successful e-government: cases from the Egyptian programme. *International Journal of Business and Social Science*, 6(1).
- Abdul-Razak, F. Z., Abu-Bakar, A., & Abdullah, W. S. (2017). How perceived effort expectancy and social influence affects the continuance of intention to use e-government: A study of a malaysian government service. *Electronic Government, An International Journal*, 13(1), 69-80.
- Abu-Shannab, E. A. (2012). Digital Government Adoption in Jordan: An Environmental Model. *International Arab Journal of E-technology*, 2(3), 129-135.
- Alateyah, S. A., Crowder, R. M., & Wills, G. B. (2013). An Exploratory Study of Proposed Factors to Adopt E-government Services: Saudi Arabia as Case Study. (*IJACSA*) *International Journal of Advanced Computer Science and Applications*, 4(11), 57-66.
- Alghamdi, S., & Beloff, N. (2014). Towards a Comprehensive Model for E-government Adoption and Utilization Analysis: The Case of Saudi Arabia. *2014 Federated Conference on Computer Science and Information Systems*, (pp. 1217-1225).
- Al-Mamari, Q., Corbitt, B., & Gekara, V. O. (2013). E-government Adoption in Oman: Motivating Factors from a Government Perspective. *Transforming Government: People, Process and Policy*, 7(2), 199-224.
- Alomari, M. K. (2014). Discovering Citizens Reaction Toward E-government: Factors in E-government Adoption. *Journal of Information Systems and Technology Management (JISTEM)*, 11(1), 05-20.
- Alsaif, M. (2013). Factors Affecting Citizens' Adoption of E-government Moderated by Socio-Cultural Values in Saudi Arabia. *ECEG2013 13th European Conference on eGovernment*, 2, pp. 578-586. Como, Italy.

- Alshehri, M., Drew, S., Alhussain, T., & Alghamdi, R. (2012). The Effects of Website Quality on Adoption of E-government Service: An Empirical Study Applying UTAUT Model using SEM. *23rd Australasian Conference on Information Systems Website Quality and E-government Service*.
- Andersen, K. V., & Henriksen, H. Z. (2006). E-government maturity models: extension of the Layne and Lee model. *Government Information Quarterly*, 23, 236-248.
- Bwalya, K. J. (2009). Factors Affecting Adoption of E-government in Zambia. *EJISDC*, 38(4), 1-13.
- Choudrie, J., & Dwivedi, Y. (2005). A Survey of Citizens' Awareness and Adoption of E-government Initiatives, The 'Government Gateway': A United Kingdom Perspective. *eGovernment Workshop '05 (eGOV05)*. London, UK.
- Colesca, S. ..., & Dobrica, L. (2008). Adoption and Use of E-government Services: The Case of Romania. *Journal of Applied Research and Technology*, 6(3), 204-217.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 318-341.
- El-Baradei, L., Shamma, H. M., & Saada, N. (2012). Examining the Marketing of E-government Services in Egypt. *International Journal of Business and Public Management*, 2(2), 12-22.
- Fath-Allah, A., Cheikhi, L., & Al-Qutaish, R. E. (2014). E-government maturity models: a comparative study. *International Journal of Software Engineering & Applications (IJSEA)*, 5(3), 71-91.
- Fink, A. (2014). *Conducting Research Literature Review: from the Internet to Paper* (4nd ed.). Thousand Oaks, CA: Sage Publications.
- Gilbert, D., Balestrini, P., & Littleboy, D. (2004). Barriers and Benefits in the Adoption of E-government. *The International Journal of Public Sector Management*, 17(4), 286-301.
- Green, B. N., Johnson, C. D., & Adams, A. (2001). Writing Narrative Literature Reviews for Peer Reviewed Journals: Secrets of Trade. *Journal of Chiropractic Medicine*, 15.
- Hu, P. J.-H., Brown, S. A., Thong, J. Y., Chan, F. K., & Tam, K. (2009). Determinants of Service Quality and Continuance Intention of Online Services: The Case of eTax. *Journal of the American Society for Information Science and Technology*, 60(2), 292-306.
- Hung, S.-Y., Tang, K.-Z., Chang, C.-M., & Ke, C.-D. (2009). User Acceptance of Intergovernmental Services: An Example of Electronic Document Management System. *Government Information Quarterly*, 26(2), 387-397.
- Komba, M. M. (2013). *Factors Influencing Access to Electronic Government Information and E-government Adoption in Selected Districts of Tanzania*. PhD Thesis, University of South Africa, Information Science.
- Kripanont, N. (2007). *Examining A Technology Acceptance Model of Internet Usage by Academics within the Business Schools*. PhD Thesis, Faculty of Business and Law, Information Systems.

- Kumar, V., Mukerji, B., Butt, I., & Persaud, A. (2007). Factors for Successful E-government Adoption: A Conceptual Framework. *The Electronic Journal of E-government*, 5(1), 63-76.
- Layne, K., & Lee, J. (2001). Developing fully functional e-government: a four stage model. *Government Information Quarterly*, 18, 122-136.
- Lin, F., Fofanah, S. S., & Liang, D. (2011). Assessing Citizen Adoption of E-government Initiatives in Gambia: A Validation of The Technology Acceptance Model in Information System Success. *Government Information Quarterly*, 28(2), 271-279.
- Mandari, H. E., & Chong, Y.-L. (2018). Gender and age differences in rural farmers' intention to use m-government services. *Electronic Government, An International Journal*, 14(3), 217-239.
- Meftahi, M., Gharleghi, B., & Samadi, B. (2015). Adoption of E-government Among Bahrain Citizens. *Asian Social Science*, 11(4), 141-149.
- Reddick, C. G., AbdelSalam, H. M., & Elkadi, H. A. (2012). Channel choice and the digital divide in e-government: The case of Egypt. *Information Technology for Development*, 18(3), 226-246.
- Rogers, E. M. (1995). *Diffusion of Innovations* (4nd ed.). New York: The Free Press.
- Rokhman, A. (2011, May). E-government Adoption in Developing Countries: The Case of Indonesia. *Journal of Emerging Trends in Computing and Information Sciences*, 2(5), 228-236.
- Srivastava, S. C., & Teo, T. (2005). Citizen trust development for e-government adoption: case of Singapora. *Pacific Asia Conference on Information Systems (PACIS)*.
- United Nations. (2014). *E-government for the future we want*. United Nations Department of Economic and Social Affairs. New York: United Nations.
- United Nations. (2016). *E-government in Support of Sustainable Development*. United Nations Department of Economic and Social Affairs. New York: United Nations.
- United Nations. (2018). *Gearing e-government to support transformation towards sustainable and resilient societies*. Department of Economic and Social Affairs. New York: United Nations.
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward A Unified View. *MIS Quarterly*, 27(3), 425-478.
- World Bank. (2016). *World Development Report 2016: Digital Dividends*. Washington, DC.

Optimization of Water Distribution Networks Using PSO Algorithm

Amira F.M.^{1,2}, A. A. Abd El-Aziz.^{1,3}, Hesham A. H¹ and Ahmed. F. A²

Abstract

In recent years there has been growing interest in water save problems. In this paper an optimizing water distribution system using an evolutionary algorithm called particle swarm optimization PSO algorithm has been introduced. A standard bench mark problem of New York water distribution system is solved using this algorithm. First simulation of the distribution system is done using simulation software; EPANET from which we integrate the head values obtained into MATLAB and the criteria for necessary head is checked. Further, the selection of different diameters and their cost estimation is done using the optimization algorithm used. The experimental result ensures the efficacy of the proposed optimization approach.

Keywords: *Swarm intelligent, Particular swarm optimization, water distributed system (WDS).*

1- Introduction

Water distribution systems are complex infrastructural entities with a combination of pipes, pumps, reservoirs, tanks etc. Each component is designed as a compromise between economical and technical considerations. It is one of the most expensive entities in any infrastructure. Simulation of these systems within a looped pipe network is a complex task involving solving the system of non-linear equations. The laying and maintenance of these water distribution systems is a high cost involving process. The growing population and migration of humans has increased the demand for water which in turn leads to the need for laying newer pipelines.

There has been tremendous research in this field using different algorithms including the recent evolutionary algorithms. The majority of those studies deals with least cost optimization however it is also essential to investigate the reliability of network design in ensuring adequate heads. Recent studies have attempted to exploit expanding computer power using these algorithms and combine these new optimization techniques with hydraulic simulation software.

¹Institute of Statistical Studies and Researches (ISSR), Cairo, Egypt.,

²Agriculture Research Center Giza, Egypt

³Information Systems, college of Computer and Information Sciences, Jouf University.

amira@capl.sci.eg, a.ahmed@cu.edu.eg, hehefny@ieee.org, afsadek@gmail.com

This project describes the development of a computer model of a water distribution system using EPANET, simulation software. The optimization algorithm is linked to EPANET via a toolkit which is used for pipe network design.

This work mainly deals with single objective optimization taking a bench mark problem such as New York water distribution system, simulating it and linking it to MATLAB. The optimization algorithm used in this is particle swarm optimization PSO algorithm, a most recent evolutionary algorithm based on the behavior of swarm and their movement in nature developed by Kennedy & Eberhart 1995[Kennedy, J. and Eberhart, R. C., 1995].

This paper contains seven sections organized as follow, first section is an introduction, section two presented an overview of standard PSO algorithm and its applications, section three talks about EPANET tool and its potential in representing and analyzes water networks, and the fourth section comes with the previous efforts in WDS. In section five the problem of New York WDS discussed and presented our approach to solve the problem. Finally, conclusion and references come in sections six and seven respectively.

2- Overview of PSO algorithm

Particle swarm optimization PSO algorithm is a new evolutionary algorithm

Introduced by Kennedy & Eberhart 1995 [Kennedy, J. and Eberhart, R. C., 1995]

, Inspired by social behavior and movement dynamics of insects, birds and fish ,
Global gradient-less stochastic search method , Suited to continuous variable problems , Performance comparable to Genetic algorithms , Has successfully been applied to a wide variety of problems (Neural Networks, Structural opt., Shape topology opt.)

2.1 PSO applications [AcarVural, et al, 2011], [Zhu, et al, 2011]

There are several applications of PSO in real world problems.

- Training of neural networks
- Identification of Parkinson's disease
- Extraction of rules from fuzzy networks
- Image recognition
- Optimization of electric power distribution networks
- Structural optimization
- Optimal shape and sizing design
- Topology optimization

- Process biochemistry
- System identification in biomechanics

2.2 General algorithm

The following algorithm used to find the global minimum of the given function with the help of the PSO algorithm, Basic algorithm as proposed by Kennedy and Eberhart (1995), see algorithm (1).

Algorithm1 : basic PSO algorithm

1. Initialize
 - (a) Set constants $k_{\max}$, c_1, c_2 .
 - (b) Randomly initialize particle positions $x_0^i \in D$ for $i = 1, \dots, p$.
 - (c) Randomly initialize particle velocities $0 < v_0^i < v_0^{\max}$ for $i = 1, \dots, p$.
 - (d) Set $k = 1$, $i = 1$.
2. Optimize
 - (a) Evaluate fitness value f_k^i using design space coordinates x_i
 - (b) If $f_k^i < f_{\text{best}}^i$ then $f_{\text{best}}^i = f_k^i$, $p_k^i = x_k^i$
 - (c) If $f_k^i < f_{\text{best}}^g$ then $f_{\text{best}}^g = f_k^i$, $p_k^g = x_k^i$
 - (d) If stopping condition is satisfied go to 3.
 - (e) Update particle velocity vector $v_{(k+1)}^i$ using (1).
 - (f) Update particle position vector $x_{(k+1)}^i$ using (2).
 - (g) Increment i . If $i > p$ then increment k , set $i = 1$.
 - (h) Go to 2(a).
3. Report results and terminate

Inertia term introduced in velocity rule:

$$v_{k+1}^i = w_k v_k^i + c_1 r_1 (p_k^i - x_k^i) + c_2 r_2 (p_k^g - x_k^i) \text{ ----(1)}$$

Position update rule remains unchanged:

$$x_{k+1}^i = x_k^i + v_{k+1}^i \text{ ----- (2)}$$

PSO algorithm flow chart

Fig.1, illustrates the flow chart of the standard PSO algorithm

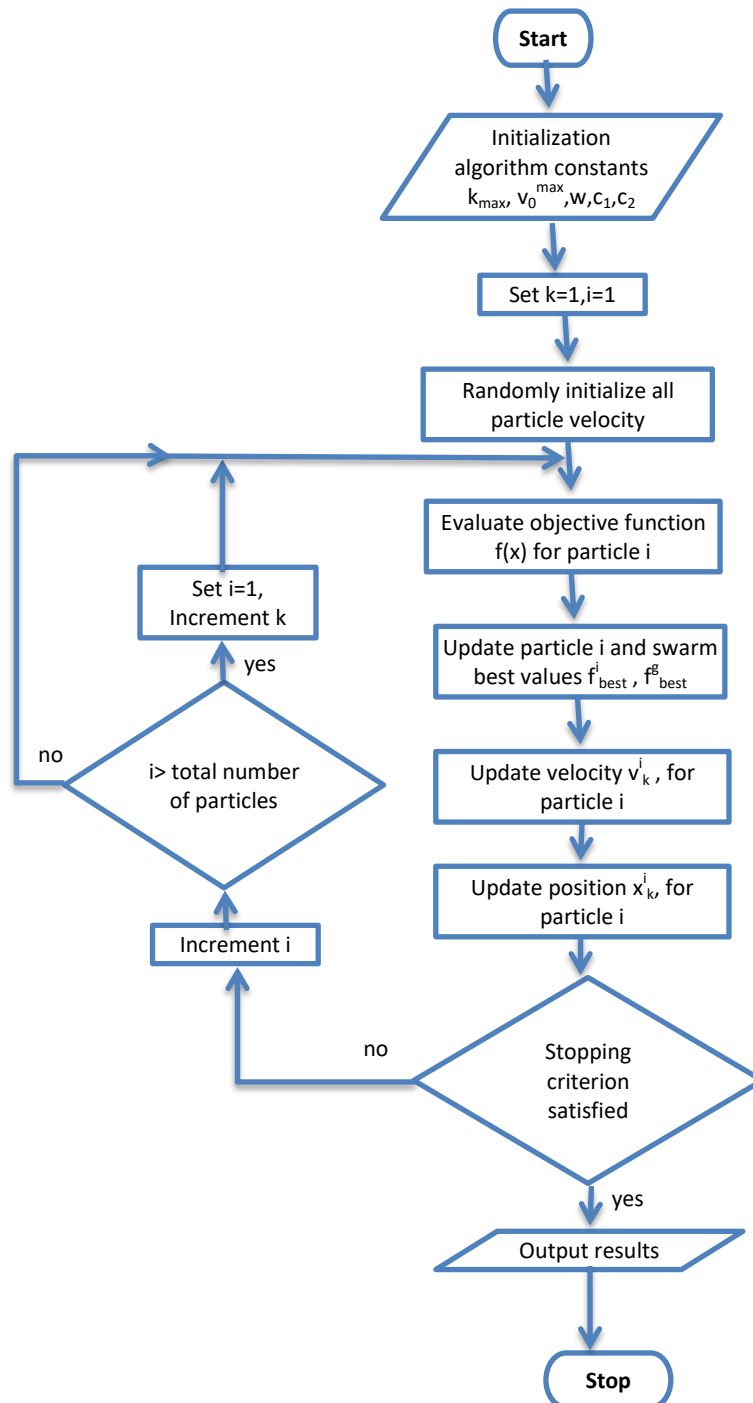


Fig.1, flowchart of the standard PSO algorithm

3- EPANET

EPANET is a computer program that performs extended period simulation of hydraulic and water quality behavior within pressurized pipe networks. A network consists of pipes, nodes (pipe junctions), pumps, valves and storage tanks or reservoirs. EPANET tracks the flow of water in each pipe, the pressure at each node, the height of water in each tank, and the concentration of a chemical species throughout the network during a simulation period comprised of multiple time steps. In addition to chemical species, water age and source tracing can also be simulated.

EPANET is designed to be a research tool for improving our understanding of the movement and fate of drinking water constituents within distribution systems. It can be used for many different kinds of applications in distribution systems analysis. Sampling program design, hydraulic model calibration, chlorine residual analysis, and consumer exposure assessment are some examples. EPANET can help assess alternative management strategies for improving water quality throughout a system. These can include: · altering source utilization within multiple source systems, · altering pumping and tank filling/emptying schedules, · use of satellite treatment, such as re-chlorination at storage tanks, · targeted pipe cleaning and replacement.

Running under Windows, EPANET provides an integrated environment for editing network input data, running hydraulic and water quality simulations, and viewing the results in a variety of formats. These include color-coded network maps, data tables, time series graphs, and contour plots [Rossman, L. A.,2000].

3.1 Hydraulic Modeling Capabilities

Full-featured and accurate hydraulic modeling is a prerequisite for doing effective water quality modeling. EPANET contains a state-of-the-art hydraulic analysis engine that includes the following capabilities [Rossman, L. A.,2000]:

- Places no limit on the size of the network that can be analyzed
- Computes friction headloss using the Hazen-Williams, DarcyWeisbach, or Chezy-Manning formulas
- Includes minor head losses for bends, fittings, etc.
- Models constant or variable speed pumps
- Computes pumping energy and cost
- Models various types of valves including shutoff, check, pressure

regulating, and flow control valves

- Allows storage tanks to have any shape (i.e., diameter can vary with height)
- Considers multiple demand categories at nodes, each with its own pattern of time variation
- Models pressure-dependent flow issuing from emitters (sprinkler heads)
- Can base system operation on both simple tank level or timer controls and on complex rule based controls.

3.2 Head loss [Rossman, L. A.,2000]

Pipes convey water from one point to another. Flow direction is from the end at higher head (potential energy per pound of water) to that at lower head. The head lost to friction associated with flow through a pipe can be expressed in a general fashion as:

$$h_L = a qb \quad (3)$$

where h_L is the head loss in feet, q is the flow in cfs, a is a resistance coefficient, and b is a flow exponent.

EPANET can use one of three popular forms of Equation 1: the Hazen-Williams formula, the Darcy-Weisbach formula, or the Chezy-Manning formula. The Hazen-Williams formula is probably the most popular head loss equation for distribution systems, the Darcy-Weisbach formula is more applicable to laminar flow and to fluids other than water, while the Chezy-Manning formula is more commonly used for open channel flow. Table 2.1 lists values of the resistance coefficients and flow exponents for each formula. Note that each formula uses a different pipe roughness coefficient that must be determined empirically. Table 2.2 lists general ranges of these coefficients for different types of new pipe materials. Be aware that a pipe's roughness coefficient can change considerably with age [Rossman, L. A.,2000].

Pipes can contain check valves in them that restrict flow to a specific direction. They can also be made to open or close at pre-set times, when tank levels fall below or above certain set-points, or when nodal pressures fall below or above certain set-points [Rossman, L. A.,2000]. Table 1 illustrates the International Headloss formulas for full pipe flow.

Table 1 Pipe Head Loss Formulas (full pipe flow)

Formula Resistance Coefficient (a) Flow Exponent (b))	
Hazen-Williams	$4.72 C^{1.85} d^{4.87} L^{1.85}$
Darcy-Weisbach	$0.0252 f(e,d,q) d^5 L^2$
Chezy-Manning	$4.66 n^2 d^{-5.33} L^2$
Where: C = Hazen-Williams roughness coefficient e = Darcy-Weisbach roughness coefficient (ft) f = friction factor (dependent on e, d, and q) d = pipe diameter (ft) L = pipe length (ft)	

4- Previous work

In recent decades scientists have been making many efforts to achieve the aims and objectives of the current research.

Miyaoka and Funabashi (1984) introduced a modelling technique and an optimal control scheme for water distribution networks. To overcome the large scale and nonlinearity of the network, a network aggregation method and a two-level control scheme were developed. The first level of the scheme decided operating points using a nonlinear optimization method, where the pressure/flow equations were solved using a high-speed technique derived from network flow theory. The second level was a feedback control around the operating points, which absorbed estimate error and small variations in consumption [Miyaoka, S. I., and Funabashi, M., 1984].

Sterling and Bargiela (1984) considered the problem of minimization of leakages due to over pressurization in a water distribution network, and presented an algorithm for computation of the optimal valve controls based on the sparse revised simplex method. To cope with the non-linearity of the system a method of iterative linearization based on the Newton-Raphson process was used. Simulation results indicated a potential for 20% reduction of the volume of leakages using optimised valve control [Sterling, M. J. H., & Bargiela, A. , 1984].

Jowitt and Xu (1990) described a successive linearization of the nonlinear network equations using the linear theory method. The resultant linear system allows linear programming techniques to minimize the leakage by determining the optimal

setting of control valves. The numerical results exhibited that the overall reduction in leakage was about 20% [Jowitt, P. W., & Xu, C. ,1990].

Chen and Powell (1990) presented an algorithm to calculate the optimal valve settings in order to reduce leakages in a water network. The optimization problem was linearized using the method of the least absolute values estimation. Based on state estimation for online monitoring of water networks, the formulation of the optimization problem lead to a linear programming problem. For such a problem, the sparse revised simplex method was employed, which showed effectiveness in speeding up solution time and enhancing numerical stability [Chen, Y. C., and Powell, R. S. 1990].

Savic and Walters (1994) presented a methodology for pressure regulation in a water distribution network encompassing the principles of evolutionary design and GAs. The optimization problem of minimizing the pressure heads was formulated with the settings of isolating valves as decision variables and minimum allowable pressures as constraints. The algorithm developed incorporates a steady-state network analysis model based on the linear theory method [Savic, D. A., and Walters, G. A. 1994].

Germanopoulos (1995) integrated the linear theory method into a linear programming technique for the excess pressure minimization problem taking into account the pressure dependent leakage term. The linear theory method was used in the iterative linearization of the nodal flow continuity constraints. The resulting linear program was solved at each iteration until convergence to an optimal solution [Germanopoulos, G. ,1995].

Vairavamoorthy and Lumbers (1998) and Alonso et al. (2000) developed an optimization method to minimize the leakage in WDSs through the most effective settings of flow reduction valves included of pressure-dependent leakage terms in network analysis. The valve setting optimization problem was formulated as a NLP problem and was solved using a sequential quadratic programming (SQP) method [Vairavamoorthy, K., & Lumbers, J., 1998].

Awad et al. (2003) presented another technique for improving the existing optimal pressure regulation and leakage minimization algorithms for supervisory water distribution networks. Self-Organized Map (SOM) and an unsupervised Artificial Neural Network (ANN) were trained with the assistance of Supervisory Control and Data Acquisition (SCADA) to classify well-regulated pressure cases for the

water distribution network based on its actual values of flow meter readings which reflect the real network water demands or consumption. After training the SOM, a simulation step was used to classify the unregulated pressure cases into the different model classes. Based on the classifications the appropriate electrical motor valves setting of the well pressure regulation events were used for the unregulated ones. Using SOM as a pre-optimization method could prevent all errors resulting from applying optimization models, save its computational time and provide an on-line pressure regulation method. The computational results showed the effectiveness of using SOM as a pre-optimization tool for regulating 74% of events within the target pressure range [Awad, H., et al., 2003].

Araujo et al. (2006; 2003) developed a model with the capacity to support decisions regarding the quantification, localization and opening adjustment of valves in a network system with the objective to minimize pressures and consequently water losses. Two operational modules were established: the first for the evaluation of an objective function to optimize the localization of valves in the network system, and the second for the opening valve adjustment in order to optimize pressures through the evaluation of an aptitude function [Araujo, et al., 2006], [Araujo, L. S., et al., 2003].

Using model of a network, Finesse automatically generates optimal network scheduling problem written in a mathematical modelling language called GAMS (Rosenthal 2007), which calls up a non-linear programming solver called CONOPT (Drud 1985) to calculate a continuous optimization solution. The optimal solution is feedback from CONOPT into Finesse for analysis and/or further processing. The optimal continuous schedule has been then discretized using an automatic algorithm coded in MATLAB [Drud, A., 1985], [Brooke, et al., 1998].

Awad et al. (2009) described a practical methodology for optimal pressure regulation in water distribution networks. The methodology could be used to design DMAs by employing PRVs to reduce excessive outlet hydraulic pressure at certain times of the day. The proposed method used GA to identify the optimal DMA boundaries and to determine the optimal type, location and setting of the PRVs. The developed optimization algorithms used both fixed and time modulated PRVs. The objective functions estimated benefits of reducing pressure on leakage, pipe burst frequency, pressure-sensitive water consumption, active leakage control effort, and energy consumption and customer contacts [Awad, H., et al., 2009]

S. Sophocleous, D. Savic, Z. Kapelan, Y. Shenand P. Sage (2017) present a systematic approach for narrowing down the search for leaks and unknown closed valves in the water distribution network. A search space reduction approach was presented that reduces the complexity of the calibration problem and contributes to earlier leak detection through inverse modelling associated with an optimization method. The approach has been formulated and applied to a real network, where the search for unknown leaks and closed valves was significantly reduced to around 10% and 30% of the network components respectively [Sophocleous, et al., 2017].

5- The proposed optimization approach

Algorithm 2 shows the pseudocode of the proposed approach

Algorithm 2 : the pseudocode of the proposed approach to solve the problem.

1. Instigate the PSO randomly as a matrix of N elements (in another words N dimensional vector). Those elements are the diameters of the required pipes.
2. Compare them with the available market diameters and round them off to the closest integer diameter.
3. Get pressure heads H at each node in the network from EPANET
4. If $H < H_{\min}$ reject this solution else move the swarm as per the algorithm to the global particle.
5. Modify the cost and leak.
6. The PSO are iterated until the end.
7. Finally choosing the best of the available solutions is done.

Fig.2 shows the flow chart of the proposed approach.

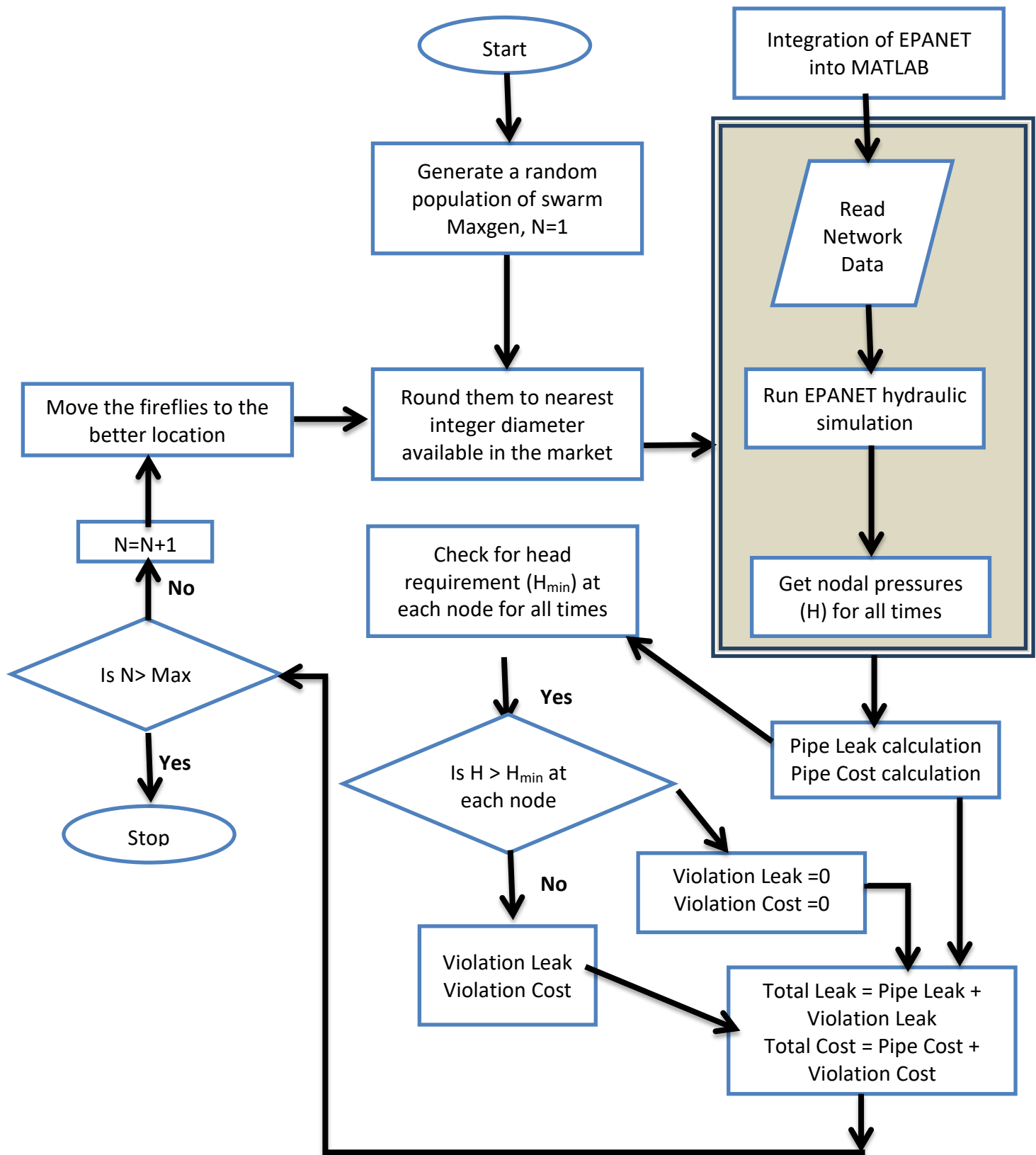


Fig.2, flowchart of the proposed approach algorithm

6- Experimental case study

New York Tunnel (NYT) benchmark network

The objective of the New York Tunnel (NYT) problem is to determine the most economically effective design of the system of tunnels that constituted the primary water distribution system of the city of New York.

All twenty-one pipes are considered for duplication. There are 15 available discrete pipe diameters (36, 48, 60, 72, 84, 96, 108, 120, 132, 144, 156, 168, 180, 192, 204) inches. The Minimum head requirement at all nodes is fixed at 255 ft except for node 16, 17 and 1 that are 260, 272.8 and 300 ft respectively.

The HeadLoss (Leakage) function is non-linear:

$$H_{\bar{y}} = 1.1 D_{\bar{y}}^{1.24} L_{\bar{y}} Q \quad \rightarrow (4)$$

Where, H is Head loss in ft./s, D is diameter in inches, L is length in ft. and Q is the network flow.

Fig.3 illustrate the case study (a) shows the network graphical design and (b) shows its simulation on EPANET

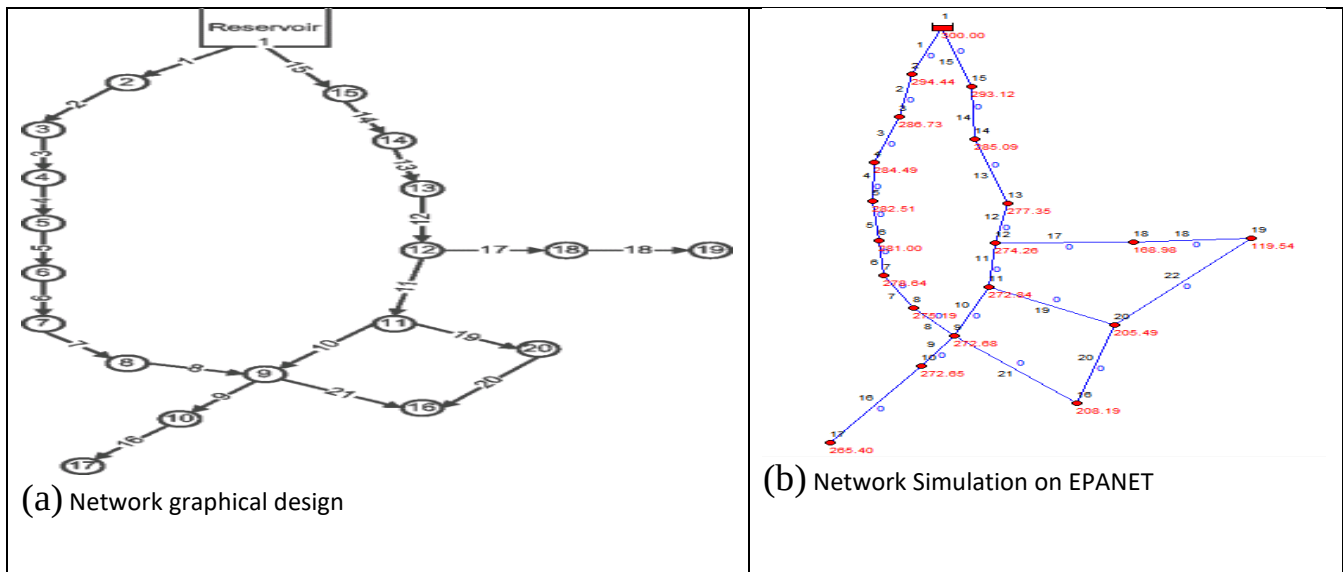


Fig 3 : Real New York water distribution network

Table 2 shows the benchmark dataset of New York example by Myrphy, Simpson and Dendy Res. Rep. No.R109, Dec. 1993

Table 2 New York example by Myrphy,Simpson and Dendy Res. Rep. No.R109, Dec. 1993																
Network Table(5.1-a) : Links					Network Table(5.1-b) : Links					Network Table(5.1-c) : Nodes						
Link ID	Length ft	Diameter in	Roughness	Flow CFS	Link ID	Velocity fps	Unit Headloss ft/Kft	Friction Factor	Status	Node ID	Base Demand CFS	Demand CFS	Head ft	Pressure psi		
Pipe 1	11600	180	100	864.79	Pipe 1	4.89	0.48	0.019	Open	Junc 2	92.4	92.40	294.44	127.58		
Pipe 2	19800	180	100	772.39	Pipe 2	4.37	0.39	0.020	Open	Junc 3	92.4	92.40	286.73	124.24		
Pipe 3	7300	180	100	679.99	Pipe 3	3.85	0.31	0.020	Open	Junc 4	88.2	88.20	284.49	123.27		
Pipe 4	8300	180	100	591.79	Pipe 4	3.35	0.24	0.020	Open	Junc 5	88.2	88.20	282.51	122.41		
Pipe 5	8600	180	100	503.59	Pipe 5	2.85	0.18	0.021	Open	Junc 6	88.2	88.20	281.00	121.76		
Pipe 6	19100	180	100	415.39	Pipe 6	2.35	0.12	0.022	Open	Junc 7	88.2	88.20	278.64	120.74		
Pipe 7	9600	132	100	327.19	Pipe 7	3.44	0.36	0.021	Open	Junc 8	88.2	88.20	275.19	119.24		
Pipe 8	12500	132	100	238.99	Pipe 8	2.51	0.20	0.022	Open	Junc 9	170.0	170.00	272.68	118.15		
Pipe 9	9600	180	100	58.50	Pipe 9	0.33	0.00	0.029	Open	Junc 10	1.0	1.00	272.65	118.14		
Pipe 10	11200	204	100	176.56	Pipe 10	0.78	0.01	0.025	Open	Junc 11	170.0	170.00	272.84	118.22		
Pipe 11	14500	204	100	511.00	Pipe 11	2.25	0.10	0.021	Open	Junc 12	117.1	117.10	274.26	118.84		
Pipe 12	12200	204	100	850.81	Pipe 12	3.75	0.25	0.020	Open	Junc 13	117.1	117.10	277.35	120.18		
Pipe 13	24100	204	100	967.91	Pipe 13	4.26	0.32	0.019	Open	Junc 14	92.4	92.40	285.09	123.53		
Pipe 14	21100	204	100	1060.31	Pipe 14	4.67	0.38	0.019	Open	Junc 15	92.4	92.40	293.12	127.01		
Pipe 15	15500	204	100	1152.71	Pipe 15	5.08	0.44	0.019	Open	Junc 16	170.0	170.00	208.19	90.21		
Pipe 16	26400	72	100	57.50	Pipe 16	2.03	0.27	0.026	Open	Junc 17	57.5	57.50	265.40	115.00		
Pipe 17	31200	72	100	222.71	Pipe 17	7.88	3.37	0.021	Open	Junc 18	117.1	117.10	168.98	73.22		
Pipe 18	24000	60	100	105.61	Pipe 18	5.38	2.06	0.023	Open	Junc 19	117.1	117.10	119.54	51.80		
Pipe 19	14400	60	100	164.44	Pipe 19	8.37	4.68	0.021	Open	Junc 20	170.0	170.00	205.49	89.04		
Pipe 20	38400	60	100	-17.05	Pipe 20	0.87	0.07	0.030	Open	Resvr 1	#N/A	-2017.50	300.00	0.00		
Pipe 21	26400	72	100	187.05	Pipe 21	6.62	2.44	0.022	Open							
Pipe 22	1000	12	100	11.49	Pipe 22	14.63	85.95	0.026	Open							

Fig.4 illustrates EPANET Real and Modified NYT simulation.,.

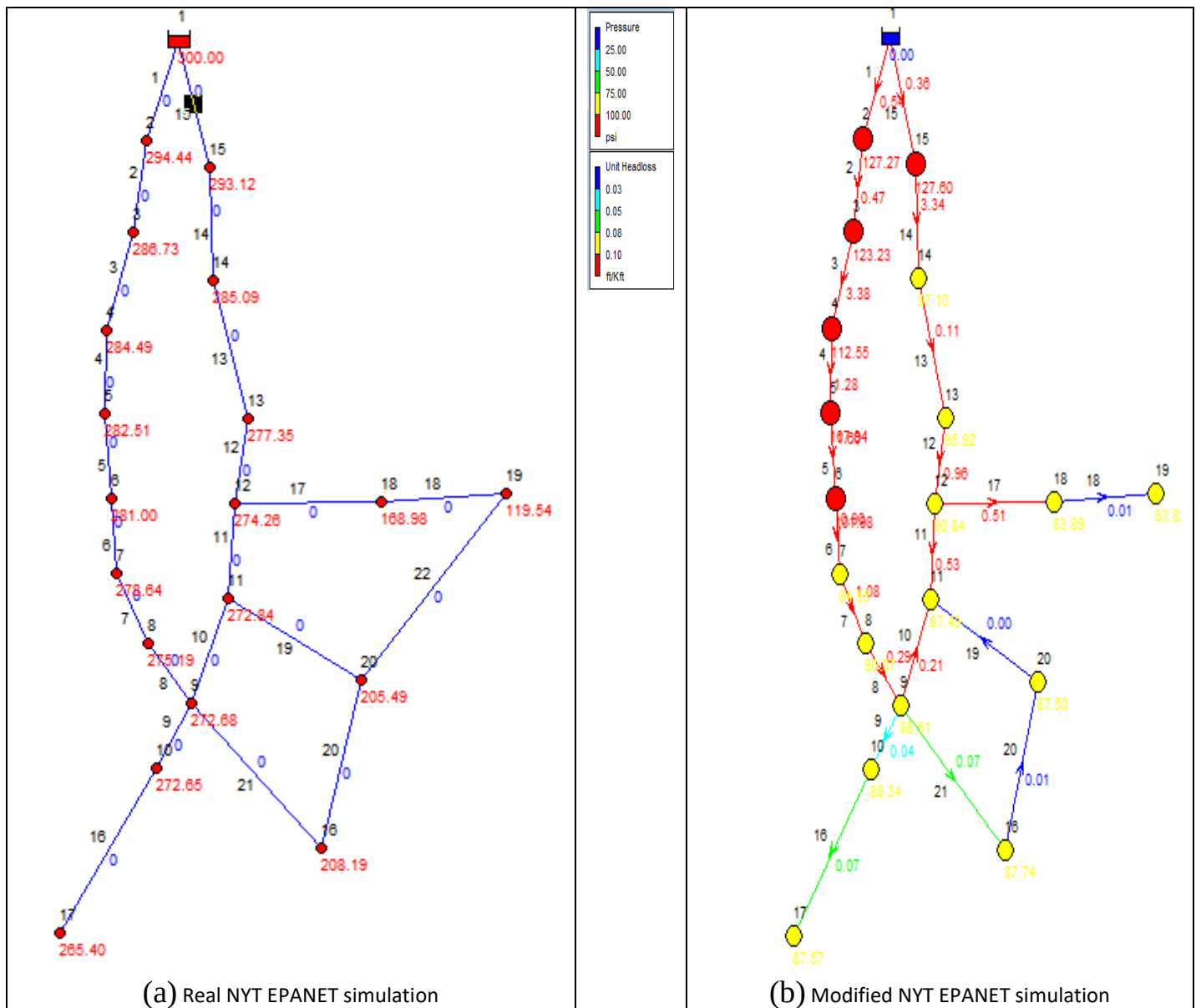


Fig. (4) : EPANET Real and Modified NYT simulation.

Table 3 illustrate pipes diameter changes after optimization

Table 3 : Pipes diameter before and after optimization		
Pipe ID	New Diameter	Old Diameter
Pipe 1	204	180
Pipe 2	204	180
Pipe 3	132	180
Pipe 4	204	180
Pipe 5	132	180
Pipe 6	156	180
Pipe 7	144	132
Pipe 8	180	132
Pipe 9	108	180
Pipe 10	84	204
Pipe 11	72	204
Pipe 12	120	204
Pipe 13	0	204
Pipe 14	108	204
Pipe 15	180	204
Pipe 16	96	72
Pipe 17	108	72
Pipe 18	204	60
Pipe 19	144	60
Pipe 20	204	60
Pipe 21	192	72
Pipe 22	-	12

Fig. 5 Explain the resulting improvement in both cost and leakage after 50 iterations. (a) Shows cost reduction, (b) shows Leakage reduction.

- Total head losses in real network = $1.66 * 10^{11}$ ft/kft
- Cost for real network = 48842510626.7009 US Dollar
- Total head losses in modified system = $4.38 * 10^{10}$ ft/kft
- Cost for modified system = 34870721521.7534 US Dollar
- Head losses reduction = 122200000000ft/kft (73.6 %)
- Cost reduction = 13971789104.9475 US Dollar (28.6%)

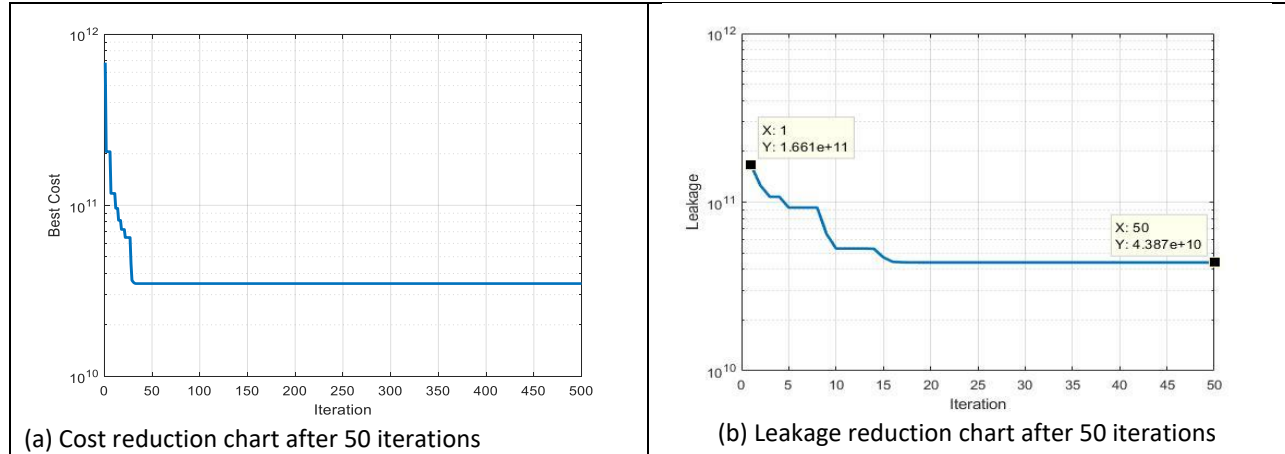


Fig.(5): Cost and Leakage reduction.

7- Conclusion

The design of water distribution networks is one of the most challenging problems optimization techniques, since it features numerous decisional variables to be considered and numerous constraints to be satisfied. It is found from the review of literature that many approaches have been developed for pipe network optimization. Most of the researchers focused on genetic algorithm to optimize the network cost.

Many other optimization techniques such as particle swarm method, ant colony optimization, artificial neural network, etc. being used for pipe network optimization data mining and ANNs are recently used for optimizing water distribution network.

In this article we comping PSO algorithm with EPANET simulation tool to optimize the network leakage. The PSO is an efficient global optimizer for continuous variable problems (structural applications), easily implemented, with very little parameters to fine-tune, Algorithm modifications improve PSO local search ability, Can accommodate constraints by using a penalty method.

The algorithm actually converges after 50–60 Iteration. The system succeed to reduce leakage for our case New York water distribution system from (1.66×10^{11} to 4.38×10^{10}) ft/kft after only 50 iteration with confederation of network requirements and available commercial pipe diameters.

Future work in this area needs to Applying the algorithm to deferent types of water networks, to achieve the national goal of water save. Also Apply the algorithm to an Egyptian water networks, to achieve the national goal of water save.

References

[AcarVural, Revna& et al., 2011]

AcarVural, Revna& Der, Ozan & Yildirim, Tülay.. "Investigation of particle swarm optimization for switching characterization of inverter design." Expert Syst. Appl.. 38. 5696-5703. 10.1016/j.eswa.2010.10.064.

[Araujo, & et al., 2003]

Araujo, L. S., Ramos, H. M., and Coelho, S. T. "Optimisation of the use of valves in a network water distribution system for leakage minimisation." Advances in Water Supply Management, C. Maksimović, D. Butler, and F. A. Memon, eds., Taylor & Francis, 97-107.

[Araujo, & et al., 2006]

Araujo, L. S., Ramos, H., and Coelho, S. T. "Pressure control for leakage minimisation in water distribution systems management." Journal of Water Resources Management , 20, 133-149.

[Awad, H., & et al.,2003]

Awad, H., Kawamura, A., Jinno, K., & Kuno, Y. "Evolutionary computing techniques for optimal pressure regulation in water distribution networks." Proceedings of Hydraulic Engineering, 47, 865-870.

[Awad, H., & et al.,2009]

Awad, H., Kapelan, Z., and Savić, D. A. "Optimal setting of time modulated pressure reducing valves in water distribution networks using genetic algorithms." the 10th international conference on Computing and Control for the Water Industry, CCWI 2009 – ‘Integrating Water Systems’ , University of Sheffield,UK.

[Brooke, A., & et al.,1998]

Brooke, A., D. Kendrick, A. Meeraus, and R. Raman. GAMS: A user's guide. GAMS Development Corporation, Washington, USA.

[Chen, Y. C., and Powell, R. S. ,1990]

Chen, Y. C., and Powell, R. S. "Pressure control through on - line monitoring." Control and Optimisation Techniques for the Water Industry, London, UK.

[Drud, A., 1985]

Drud, A. Conopt: A grg code for large sparse dynamic non-linear optimisation problems. Mathematical Programming 31, 153– 191.

[Germanopoulos, G. ,1995]

Germanopoulos, G. Valve control regulation for reducing leakage. In Improving efficiency and reliability in water distribution systems (pp. 165-188). Springer, Dordrecht.

[Jowitt, P. W., & Xu, C. ,1990]

Jowitt, P. W., & Xu, C. "Optimal valve control in water-distribution networks." *Journal of Water Resources Planning and Management*, 116(4), 455-472.

[Kennedy, J. and Eberhart, R. C., 1995]

Kennedy, J. and Eberhart, R. C. "Particle swarm optimization". *Proc. IEEE Int'l. Conf. on Neural Networks, IV*, 1942–1948. Piscataway, NJ: IEEE ServiceCenter.

[Miyaoka, et al. ,1984]

Miyaoka, S. I., & Funabashi, M. "Optimal control of water distribution systems by network flow theory". *IEEE transactions on automatic control*, 29(4), 303-311.

[Rossman, L. A.,2000]

Rossman, L. A. EPANET 2: user's manual.

[Savic, D. A., and Walters, G. A. ,1994]

Savic, D. A., and Walters, G. A. "An evolution program for pressure regulation in water distribution networks." 94/15, Centre for Systems and Control Engineering, University of Exeter, Exeter.

[Sophocleous, & et al., 2017]

Sophocleous, Sophocles & Savic, Dragan & Kapelan, Zoran & Shen, Y & Sage, P. A model pre-processing approach for improving calibration-based leakage detection using a genetic algorithm. 10.15131/shef.data.5364604.v1.

[Sterling and et al., 1984]

Sterling, M. J. H., & Bargiela, A. "Leakage reduction by optimised control of valves in water networks." *Transactions of the Institute of Measurement and Control*, 6(6), 293-298.

[Vairavamoorthy, K., & Lumbers, J., 1998]

Vairavamoorthy, K., & Lumbers, J. Leakage reduction in water distribution systems: optimal valve control. *Journal of hydraulic Engineering*, 124(11), 1146-1154.

[Zhu, H. and et al.,2011]

Zhu, H., Wang, Y., Wang, K., & Chen, Y. (2011). "Particle Swarm Optimization (PSO) for the constrained portfolio optimization problem." *Expert Systems with Applications*, 38(8), 10161-10169.

تحسين شبكات توزيع المياه باستخدام خوارزمية PSO
أميرة فرح محمد^{1,2} ، عبد العزيز أحمد عبد العزيز^{3,1} ، هشام احمد حفنى² ، أحمد فاروق الصادق²

الملخص

في السنوات الأخيرة كان هناك اهتمام متزايد بمشكلة توفير المياه. في هذا البحث تم إدخال نظام محسن لشبكات المياه باستخدام الخوارزمية التطورية PSO. تم حل مشكلة في شبكة توزيع المياه في نيويورك باستخدام هذه الخوارزمية. يتم أولاً محاكاة لنظام التوزيع باستخدام برنامج المحاكاة EPANET2 الذي نقوم من خلاله بدمج قيم التسرب عند الرؤوس التي تم الحصول عليها في MATLAB ويتم التحقق من معايير الرؤوس الضرورية. ثم يتم اختيار أقطار مختلفة للتحكم في معدل السريان وضغط الشبكة وتقدير تكلفة الشبكة باستخدام خوارزمية التحسين.

¹ معهد الدراسات والبحوث الإحصائية جامعة القاهرة ، مصر

² مركز البحوث الزراعية ، الجيزة ، مصر

³ قسم نظم المعلومات - كلية علوم الحاسب والمعلومات- جامعة الجوف

amira@capl.sci.eg, a.ahmed@cu.edu.eg, hehefny@ieee.org, afsadek@gmail.com

Deep Learning Based Diabetic Retinopathy Diagnosis: A Survey

Ahmed M. Abdo, Ahmed H. Asad and Hesham A. Hefny

1 Abstract

Diabetic retinopathy (DR) is one of the main retinal abnormalities which is asymptomatic and is the main cause of vision loss in diabetic patients. Automation is a prerequisite to improve health care efficiency by using a computer-aided diagnosis systems to facilitate the doctor and also decrease the diagnosis time. Using deep learning algorithms, in particular convolution networks, have rapidly become a methodology of choice for analyzing medical images. This paper reviews major deep learning techniques and model structures pertinent to retinal image classification. We survey the use of deep neural networks for image classification and detection of diabetic retinal diseases in patients.

Keywords: Deep Learning, Convolutional Neural Networks, Diabetes, Diabetic Retinopathy, Image Classification.

2 Introduction

Diabetes is one of the major diseases being faced by the world today. World Health Organization(WHO) survey estimated that 2.8% people suffered from diabetes in 2000, and this percentage would increase to 4.4% in 2030 [Wild et al., 2004]. Diabetes is becoming common nowadays in people due to physical inactivity, obesity, and aging population. Diabetic retinopathy (DR) is a secondary micro vascular complication of both type 1 and type 2 diabetes, the prevalence of which strongly correlates to both the duration of diabetes and the level of glycemic control as evidenced by diabetes control and complication trial and UK prospective diabetes study [UK et al., 1998, Lee et al., 2015].

DR is the most frequent cause of new cases of blindness among the adults aged 20–64 years in the developed countries [Crawford et al., 2009] . It is classified into non-proliferative DR (NPDR) and proliferative DR (PDR) stages Figure 1.

The earliest clinical sign of NPDR includes micro aneurysms which appear as small red dots in the superficial retinal layers and cause the retinal hemorrhaging. The dot and blot hemorrhages occur as micro aneurysms rupture in the deeper layers of the retina such as the inner nuclear and the outer plexiform layer. This is followed by the flame-shaped hemorrhages which occur in more superficial layers of the retina. Later as the disease progresses, the cotton-wool spots, venous beading, and the intra retinal micro vascular abnormalities develop, which are the hallmarks of the progressive capillary perfusion [Bhavsar and Bhavsar, 2016] . Neovascularization on the surface of the retina and the optic disc in conjunction with further retinal ischemia signifies the presence of the PDR [Crawford et al., 2009, Mohamed et al., 2007], Figure 2 shows how a healthy retinal image looks like compared to the one in Figure 1.

The current detection process of DR is manual, costly, and time-consuming and requires specialized skills to operate the equipment. Therefore, a new paradigm for the care of these patients is needed. Periodic sequential retinal photos are rapidly becoming the standard of care for monitoring various ocular conditions such as glaucoma, diabetic and hypertensive retinopathy, and macular degeneration. The computer-aided diagnosis systems based on image processing are becoming common these days to facilitate the doctor and reduce the diagnostic time.

In order to reduce the workload of human interpretation, and therefore streamline retinal pathology screening, various image analysis and detection algorithms have been developed over the last few decades. Solutions were trained to detect lesions using manual segmentation for supervision [Winder et al., 2009, Abramoff et al., 2010] : this is what we call computer-aided detection algorithms. Based on the detected lesions, other algorithms were trained to detect pathologies [Abramoff et al., 2010] : this is what we call computer-aided diagnosis algorithms. In recent years, new algorithms were designed to detect pathologies directly, using diagnoses only for supervision:

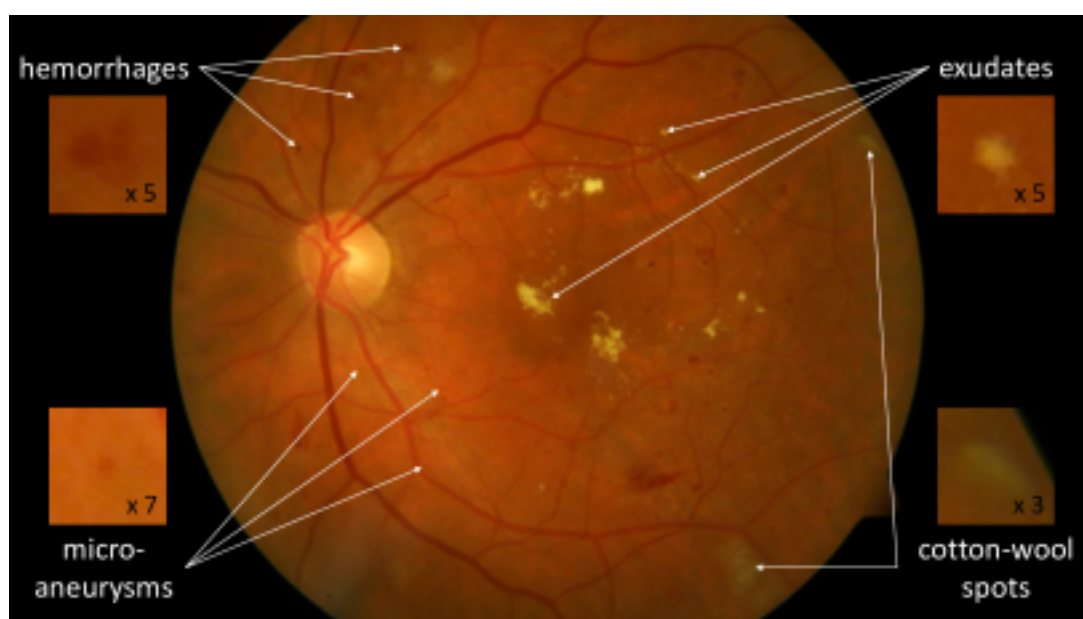


Figure 1: Most common lesions caused by diabetic retinopathy.

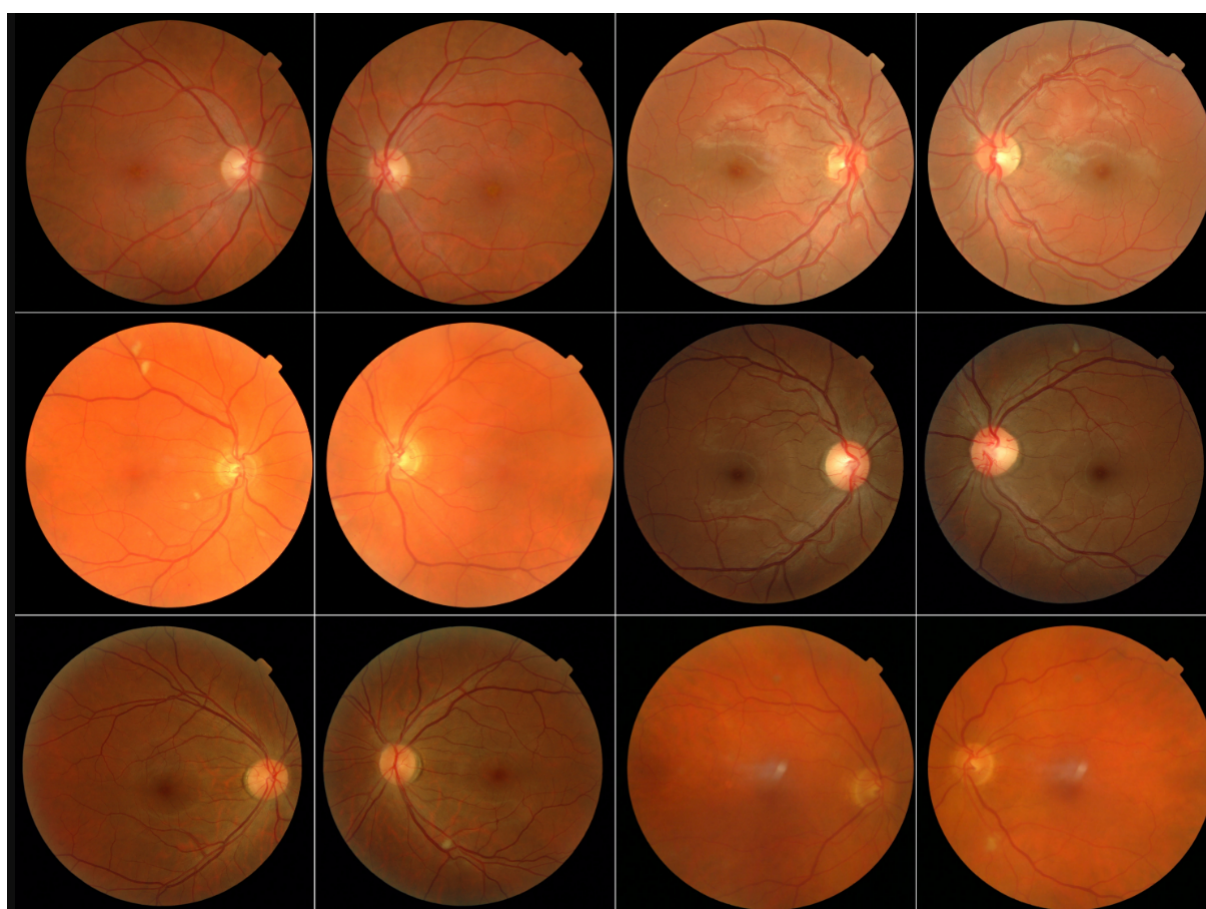


Figure 2: Retinal fundus image samples

these algorithms are based on multiple-instance learning [Quelleg et al., 2017, Manivannan et al., 2017] or deep learning [LeCun et al., 2015]. Because manual segmentation are not needed, such algorithms can be trained with much larger datasets, such as anonymized archives of examination records. A few multiple-instance learning algorithms, supervised at the image level, can already detect lesions [Melendez et al., 2015, Quelleg et al., 2016].

The motivation for our review was to offer a comprehensive overview of (almost) all fields in medical imaging related to Diabetic Retinopathy, both from an application and a methodology-drive perspective. This also includes overview tables of all publications which readers can use to quickly assess the field. Last, we leveraged our own experience with the application of deep learning methods to medical image analysis to provide readers with a dedicated discussion section covering the state-of-the-art, open challenges and overview of research directions and technologies that will become important in the future.

Summarizing, with this survey we aim to:

- Show that deep learning techniques have permeated the entire field of retinal image analysis;
- Identify the challenges for successful application of deep learning to retinal imaging tasks;
- Highlight specific contributions which solve or circumvent these challenges.

The rest of this survey as structured as followed. In Section 3 we introduce the main deep learning techniques that have been used for retinal image analysis and that are referred to throughout the survey. Section 4 describes the contributions of deep learning to canonical tasks in Diabetic Retinopathy detection and detailed reviews to state of the art in deep learning for retinal image analysis.

3 Overview of deep learning methods

The goal of this section is to provide a formal introduction and definition of the deep learning concepts, techniques and architectures that we found in the medical image analysis papers surveyed in this work.

3.1 Learning algorithms

Machine learning methods are generally divided into supervised and unsupervised learning algorithms, although there are many nuances. In supervised learning, a model is presented with a dataset $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\}_{n=1}^N$ of input features $\mathbf{x}$ and label $\mathbf{y}$ pairs, where $\mathbf{y}$ typically represents an instance of a fixed set of classes. In the case of regression tasks $\mathbf{y}$ can also be a vector with continuous values. Supervised training typically amounts to finding model parameters Θ that best predict the data based on a loss function $L(\mathbf{y}, \hat{\mathbf{y}})$. Here $\hat{\mathbf{y}}$ denotes the output of the model obtained by feeding a data point $\mathbf{x}$ to the function $f(\mathbf{x}; \Theta)$ that represents the model.

Unsupervised learning algorithms process data without labels and are trained to find patterns, such as latent subspace. Examples of traditional unsupervised learning algorithms are principal component analysis and clustering methods. Unsupervised training can be performed under many different loss functions. One example is reconstruction loss $L(\mathbf{x}, \hat{\mathbf{x}})$ where the model has to learn to reconstruct its input, often through a lower-dimensional or noisy representation.

3.2 Neural Networks

Neural networks are a type of learning algorithm Figure 4 which forms the basis of most deep learning methods. A neural network consists of neurons or units with some activation a and parameters $\Theta = \{\mathcal{W}, \mathcal{B}\}$, where $\mathcal{W}$ is a set of weights and $\mathcal{B}$ a set of biases. The activation represents a linear combination of the input $\mathbf{x}$ to the neuron and the parameters, followed by an element-wise non-linearity $\sigma(\cdot)$, referred to as a transfer function:

$$a = \sigma(\mathbf{w}^T \mathbf{x} + b).$$

Typical transfer functions for traditional neural networks are the sigmoid and hyperbolic tangent function. The multi-layered perceptrons (MLP), the most well-known of the traditional neural networks, have several layers of these transformations:

$$f(\mathbf{x}; \Theta) = \sigma(\mathbf{W}^T \sigma(\mathbf{W}^T \dots \sigma(\mathbf{W}^T \mathbf{x} + b)) + b).$$

Here, $\mathbf{W}$ is a matrix comprising of columns $\mathbf{w}_k$, associated with activation K in the output. Layers in between the input and output are often referred to as 'hidden' layers. When a neural network contains multiple hidden layers it is typically considered a 'deep' neural network, hence the term 'deep learning'.

At the final layer of the network the activations are mapped to a distribution over classes $P(y|\mathbf{x}; \Theta)$ through a softmax function:

$$P(y|\mathbf{x}; \Theta) = \text{softmax}(\mathbf{x}; \Theta) = \frac{e^{\mathbf{w}_i^T \mathbf{x} + b_i}}{\sum_{k=1}^K e^{\mathbf{w}_k^T \mathbf{x} + b_k}},$$

where $\mathbf{w}_i$ indicates the weight vector leading to the output node associated with class i . A schematic representation of three-layer MLP is shown in Figure 3.

Maximum likelihood with stochastic gradient descent is currently the most popular method to fit parameters Θ to a dataset $\mathcal{D}$. In stochastic gradient descent a small subset of the data, a mini-batch, is used for each gradient update instead of the full data set. Optimizing maximum likelihood in practice amounts to minimizing the negative log-likelihood:

$$\arg \min_{\Theta} - \sum_{n=1}^N \log [P(y_n|\mathbf{x}_n; \Theta)].$$

This results in the binary cross-entropy loss for two-class problems and the categorical cross-entropy for multi-class tasks. A downside of this approach is that it typically does not optimize the quantity we are interested in directly, such as area under the receiver-operating characteristic (ROC) curve or common evaluation measures for segmentation, such as the Dice coefficient.

For a long time, deep neural networks (DNN) were considered hard to train efficiently. They only gained popularity in 2006 [Bengio et al., 2007, Hinton and Salakhutdinov, 2006, Hinton et al., 2006] when it was shown that training DNNs layer-by-layer in an unsupervised manner (pre-training), followed by supervised fine-tuning of the stacked network, could result in good performance. Two popular architectures trained in such a way are stacked auto-encoders (SAEs) and deep belief networks (DBNs). However, these techniques are rather complex and require a significant amount of engineering to generate satisfactory results.

Currently, the most popular models are trained end-to-end in a supervised fashion, greatly simplifying the training process. The most popular architectures are convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs are currently most widely used in (medical) image analysis, although RNNs are gaining popularity. The following sections will give a brief overview of convolutional neural networks (CNNs) as it is our main focus on this survey.

3.3 Convolutional Neural Networks (CNNs)

There are two key differences between MLPs and CNNs. First, in CNNs weights in the network are shared in such a way that it the network performs convolution operations on images. This way, the model does not need to learn separate detectors for the same object occurring at different positions in an image, making the network equivariant with respect to translations of the input. It also drastically reduces the amount of parameters (i.e. the number of weights no longer depends on the size of the input image) that need to be learned. An example of a 1D CNN is shown in Figure 3.

At each layer, the input image is convolved with a set of K kernels $\mathcal{W} = \{\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_K\}$ and added biases $\mathcal{B} = \{b_1, \dots, b_K\}$, each generating a new feature map $\mathbf{X}_k$. These features are subjected to an element-wise non-linear transform $\sigma(\cdot)$ and the same process is repeated for every convolutional layer l :

$$\mathbf{X}_k^l = \sigma(\mathbf{W}_k^{l-1} * \mathbf{X}^{l-1} + b_k^{l-1}).$$

The second key difference between CNNs and MLPs, is the typical incorporation of pooling layers in CNNs, where pixel values of neighborhoods are aggregated using a permutation invariant function, typically the max or mean operation. This induces a certain amount of translation invariance and again reduces the amount of parameters in the network. At the end of the convolutional stream of the network, fully-connected layers (i.e. regular neural network layers) are usually added, where weights are no longer shared. Similar to MLPs, a distribution over classes is generated by feeding the activations in the final layer through a softmax function and the network is trained using maximum likelihood.

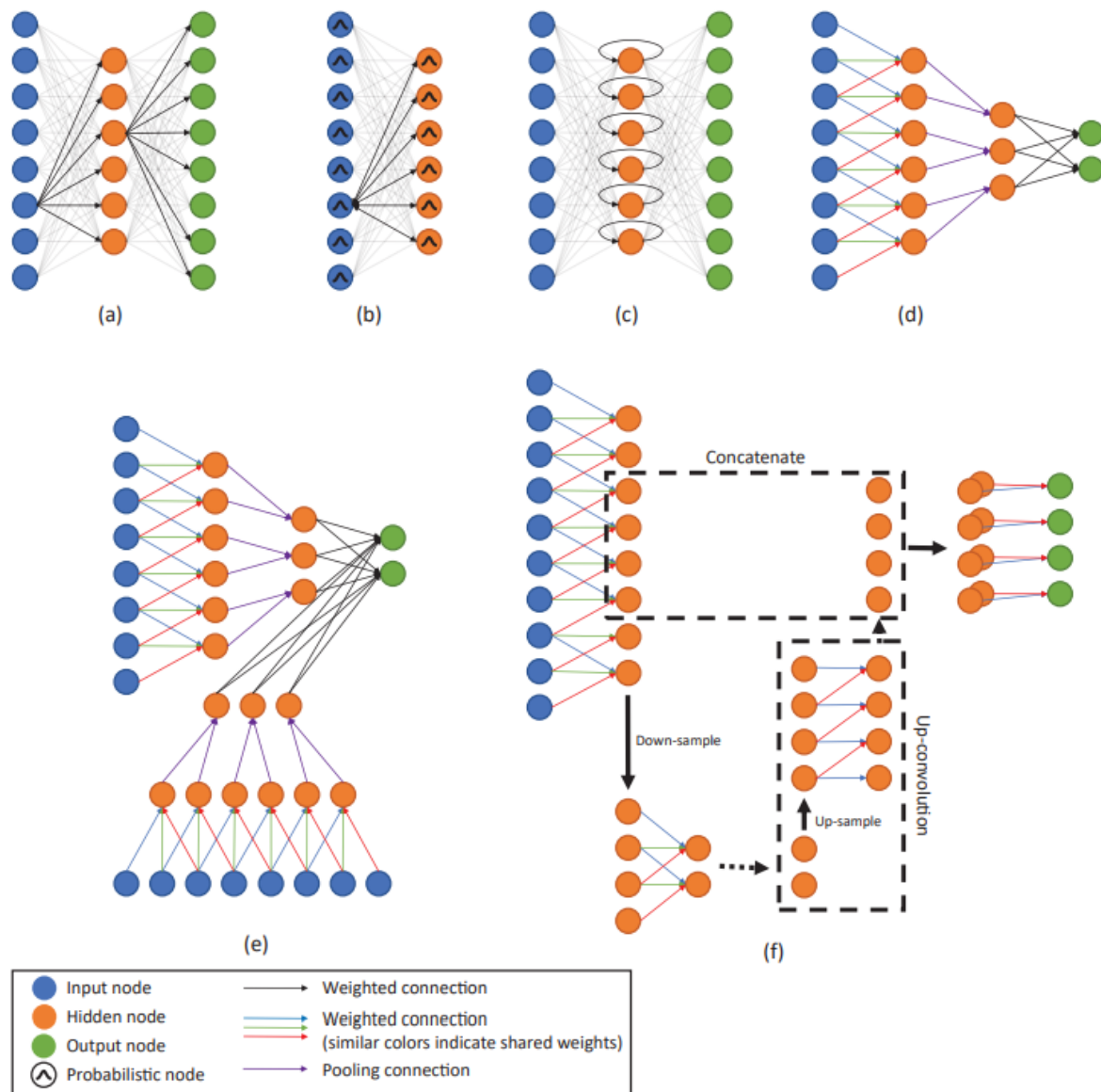


Figure 3: Architectures used in medical imaging. a) Auto-encoder, b) restricted Boltzmann, c) recurrent neural network, d) convolutional neural network, e) multi-stream convolutional neural network, f) U-net (with a single downsampling stage).

3.4 Hardware and Software

One of the main contributors to steep rise of deep learning has been the widespread availability of GPU and GPU-computing libraries (CUDA, OpenCL). GPUs are highly parallel computing engines, which have an order of magnitude more execution threads than central processing units (CPUs). With current hardware, deep learning on GPUs is typically 10 to 30 times faster than on CPUs.

Next to hardware, the other driving force behind the popularity of deep learning methods is the wide availability of open source software packages. These libraries provide efficient GPU implementations of important operations in neural networks, such as convolutions; allowing the user to implement ideas at a high level rather than worrying about low-level efficient implementations. At the time of writing, the most popular packages were (in alphabetical order):

- **Caffe** [Jia et al., 2014] . Provides C++ and Python interfaces, developed by graduate students at UC Berkeley.

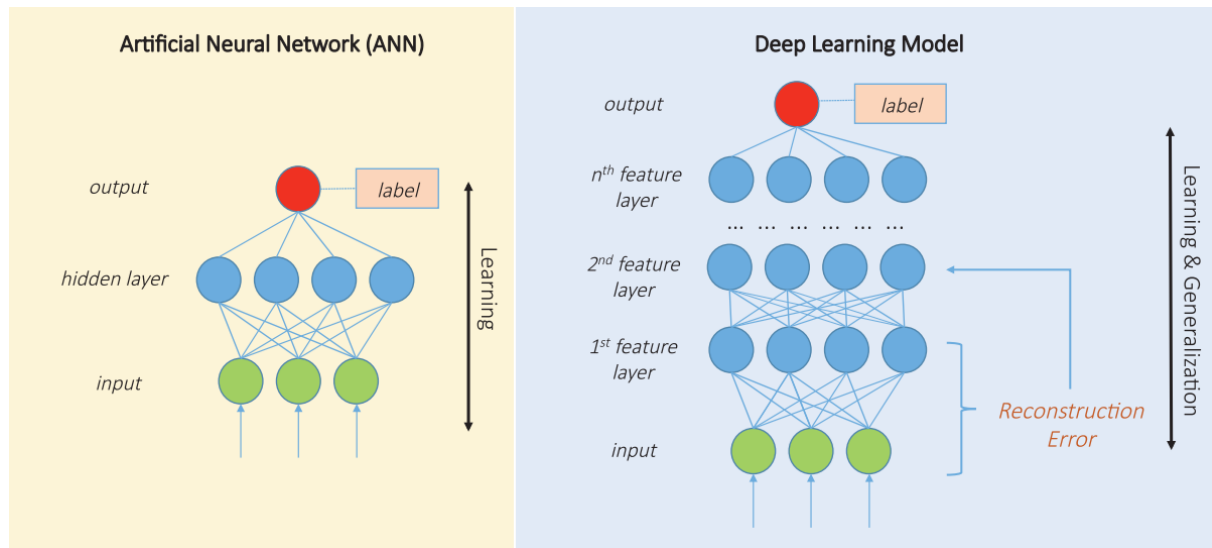


Figure 4: Comparison between ANNs and deep architectures. While ANNs are usually composed by three layers and one transformation toward the final outputs, deeplearning architectures are constituted by several layers of neural networks. Layer-wise unsupervised pre-training allows deep networks to be tuned efficiently and to extract deep structure from inputs to serve as higher-level features that are used to obtain better predictions.

- **Tensorflow** [Abadi et al., 2016] . Provides C++ and Python interfaces, developed by Google and is used by Google research.
- **Theano** [Bastien et al., 2012] . Provides a Python interface, developed by MILA lab in Montreal.
- **Torch** [Collobert et al., 2011] . Provides a Lua interface and is used by, among others, Facebook AI research.
- **PyTorch** Provides a Python interfaces for Torch and is very popular among researchers.

There are third-party packages written on top of one or more of these frameworks, such as Lasagne (<https://github.com/Lasagne/Lasagne>) or Keras (<https://keras.io/>). It goes beyond the scope of this paper to discuss all these packages in detail.

4 Deep Learning Based Diabetic Retinopathy Detection

Ophthalmic imaging has developed rapidly over the past years, but only recently are deep learning algorithms being applied to eye image understanding. As summarized in Table 1, most works employ simple CNNs for the analysis of color fundus imaging (CFI). A wide variety of applications are addressed: detection of retinal abnormalities, diagnosis of eye diseases, and image quality assessment.

For algorithm development, macula-centered retinal fundus images were retrospectively obtained from EyePACS in the United States among patients presenting for diabetic retinopathy screening. All images were de-identified according to Health Insurance Portability and Accountability Act Safe Harbor prior to transfer to study investigators. Ethics review and institutional review board exemption was obtained using Quorum Review IRB.

In 2015, a machine learning competition was organized with the goal to design an automated system for grading the severity of diabetic retinopathy in images. Images were trained and evaluated in a publicly-available dataset of almost 90,000 images provided by EyePACS [Cuadros and Bresnick, 2009], a free platform for diabetic retinopathy screening. As usual in recent pattern recognition competitions [Russakovsky et al., 2015], the top-ranking solutions all relied on deep learning. More precisely, they relied on ConvNets.

EyePACS dataset contains over 35,000 color fundus images were provided to train algorithms to predict the severity of disease in 53,000 test images. The majority of the teams that entered the competition; applied deep learning which achieved performance above that of humans, all using end-to-end Convolution neural networks (CNN).

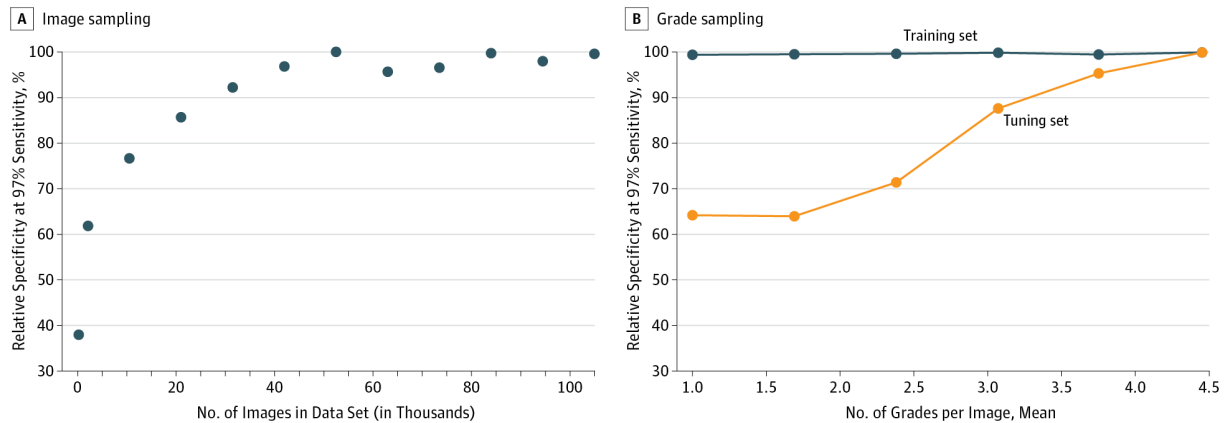


Figure 5: Model Performance for Detection of Diabetic Retinopathy

In a Kaggle machine-learning competition, [EyePACS, 2015] deep learning was used to predict the diabetic retinopathy grade only (no diabetic macular edema prediction). The winning entry had a performance that was comparable with an ophthalmologist grading the wrong eye of the same patient and higher than the agreement demonstrated between general physicians and ophthalmologists [Graham and van der Maaten, 2017].

Solanki et al. [Solanki et al., 2015] reported a sensitivity of 93.8% at a specificity of 72.2% on the same data set.

Advantages:

- Commercially available system and was being used by Eyenuk Inc, Woodland Hills, CA, USA.
- Automated screening tool.

Disadvantages:

- Trained and evaluated in a small dataset (Messidor2).
- High error rate in the moderate Non-proliferative DR.

Abràmoff et al. [Abramoff et al., 2016] reported a sensitivity of 96.8% at a specificity of 59.4% for detecting referable diabetic retinopathy on the publicly available Messidor-2 data set. [Decencièrè et al., 2014]

Advantages:

- End to End CNN.
- Outperforms traditional method.
- Three board certified retinal specialists independently graded all images used in training

Disadvantages:

- Underpowered to determine the detection performance of isolated Proliferative DR without Macular edema.
- Diabetic retinopathy and Macular edema prevalence and severity was underestimated in this study.

Burlina et al. [Burlina et al., 2016] use the OverFeat features, derived from a ConvNet trained on the very large, general-purpose ImageNet dataset: these features are used to train a linear support-vector machine (LSVM) and achieved between nearly 92% and 95% accuracy.

Advantages:

- Tests using over 5600 images from the NIH AREDS dataset.
- Used transference learning to speed up the training process.

Disadvantages:

- Semi automated approach.
- Manually extracted the OverFeat features from the CNN by tapping into the network layer 19.
- focused only on Age-related macular generation.

Van Grinsven et al. [van Grinsven et al., 2016] presented a method to substantially speed-up the time-consuming training process of convolutional neural networks with a selective sampling strategy, named SeS, embedded in the training procedure with an accuracy of 0.972.

Advantages:

- Used a small number of epochs to train the Neural network.
- Focusing the training effort on the image samples that is not well represented in the training set which pose greater difficulty.

Disadvantages:

- Did not work well in mid and low quality images.
- Hyper-parameters was not performed in this study, The depth of the network, i.e. the number of layers, and the number of kernels per layer, as well as the use of fully connected layers should be more thoroughly investigated.

Gulshan et al. [Gulshan et al., 2016] performed a thorough analysis of the performance of a Google Inception v3 network for diabetic retinopathy detection Figure 5, showing performance comparable to a panel of seven certified ophthalmologists. The presence of referable DR is detected at the image level using an ensemble of ConvNets to generate an algorithm with 97.5% sensitivity and 93.4% specificity.

Advantages:

- Used a large data set with multiple grades per image.

Disadvantages:

- The algorithm is only identifying diabetic retinopathy and diabetic macular edema missing non-diabetic retinopathy that it was not trained to identify.

Table 1: Color fundus images: detection of abnormalities and diseases

[Solanki et al., 2015]	EyeArt: automated, high-throughput, image analysis for diabetic retinopathy screening.
[Abramoff et al., 2016]	Diabetic retinopathy detection; end-to-end CNN, outperforms traditional method, evaluated on a public dataset.
[Burlina et al., 2016]	Age-related macular degeneration detection; uses overfeat pretrained network for feature extraction.
[van Grinsven et al., 2016]	Hemorrhage detection; CNN dynamically trained using selective data sampling to perform hard negative mining.
[Gulshan et al., 2016]	Diabetic retinopathy detection; Inception network, performance comparable to a panel of seven certified ophthalmologists.
[Prentasić and Lončarić, 2016]	Hard exudate detection; end-to-end CNN combined with the outputs of traditional classifiers for detection of landmarks.
[Colas et al., 2016]	Deep learning approach for diabetic retinopathy screening.
[Arunkumar and Karthigaikumar, 2017]	Multi-retinal disease classification by reduced deep learning features.
[Suriyal et al., 2018]	Mobile assisted diabetic retinopathy detection using deep neural network.
[Desbiens et al., 2018]	Deep Annotated Learning, Harmonic Descriptors and Automated Diabetic Retinopathy Detection
[Krause et al., 2018]	Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy.

Prentasić et al. [Prentasić and Lončarić, 2016] Used the detection of exudates in color fundus photographs for detection of diabetic retinopathy. They used deep convolutional neural networks for exudate detection and in order to incorporate high level anatomical knowledge about exudate locations, output of the CNN is combined with the output of the optic disc detection and vessel detection procedures. Resulted in maximum F1 measure of 0.78.

Advantages:

- Used a deep convolutional neural networks to segment exudates in color fundus photographs.

Disadvantages:

- The algorithms took 200,000 iterations to converge.
- Evaluated on a small data set(DRiDB).
- No preprocessing and postprocessing steps have been made which would improve the final result.

Colas et al. [Colas et al., 2016] also use deep learning techniques for detecting referable DR, lesions are first detected (at the pixel level) and then DR severity is graded (at the image level). The algorithm achieves an area under the receiver operating characteristic curve AUROC of 0.946 with 96.2% sensitivity and 66.6% specificity.

Advantages:

- Algorithm was trained on over 70,000 labeled retinal images.
- With a multi graded by ophthalmologists as follows: 0 (no retinopathy), 1 (mild non proliferative DR), 2 (moderate non proliferative DR), 3 (severe non proliferative DR) and 4 (proliferative retinopathy).

Disadvantages:

- Trained on a private dataset.

Arunkumar et al. [Arunkumar and Karthigaikumar, 2017] proposed an algorithm for differentiating multiple pathologies: AMD, DR, macula pucker, retinoblastoma, retinal detachment and retinitis pigmentosa. Similarly to the AMD detector, image features are extracted with a ConvNet and a multi-class SVM is used to differentiate the various pathologies resulted in an accuracy of 0.093.

Advantages:

- Multi screening system that is able to diagnosing retina for diseases such as age-related molecular degeneration, diabetic retinopathy, macular bunker, retinoblastoma, retinal detachment, and retinitis pigmentosa.
- Used an ensemble method of CNN and SVM for classification.

Disadvantages:

- Build using two datasets (DRIVE, REVIEW) which have a low number of samples for train and test.
- Despite the ability of detecting more than one disease affected the overall accuracy of the model.

Desbiens J et al. [Desbiens et al., 2018] uses two-stage process involving a clinically relevant pre-processing stage for localizing/annotating micro – aneurysms, hemorrhage and other lesions, Then using a pre-trained resnet fined tuned for Diabetic Retinopathy (DR), to essentially perform a binary classification between certain DR conditions. The algorithm implements the above methodology on two datasets: privately held dataset and publicly available messidor-2 dataset. Even though the pre-processing step presents clinical relevance, the paper lacks novelty in the area of deep learning methods with only a fine tuned resnet model been used for binary DR classification.

Advantages:

- Used a variety of cameras to shoot the pictures that will be used in training.
- Used the radial symmetry map to detect dark and bright lesions.

Disadvantages:

- alot of pre-processed and pre-digested image has been done and it was not clear on the paper.

Recently Suriyal S et al. [Suriyal et al., 2018] used a tensorflow deep neural network architecture that is trained and tested on 16,798 fundus images. These images are preprocessed to remove noise and prepare them to be fed into neural network. Preprocessing steps involve averaging all the images using a 5×5 filter and 28 convolutional layers and after each layer there is batchnorm and ReLU nonlinear function except at the final layer. The final accuracy of the model is 73.3%. and the model is optimized to work on mobile devices and does not require Internet connection to run.

Advantages:

- Mobile friendly neural network which can be used in the future to be embedded in an IOT device.

Disadvantages:

- Optimizing the neural network to make it mobile friendly affected the performance of the network.

Finally Krause J et al. [Krause et al., 2018] used an ensemble of ConvNets which had the same architecture as Gulshan et al. [Gulshan et al., 2016] used but instead of using 299×299 image dimensions they used 779×779 image dimensions and the other significant change was in the last layer which predicts a 5-class rating for DR (Gulshan et al. predicts binary referables). The final accuracy of the model is 98.3% Figure 6.

Advantages:

- State of the art on detecting Diabetic Retinopathy.
- Multi graded Retinal images.
- Ensemble method that enhanced the final result.

Disadvantages:

- Grading of DR is a complex task and agreement between clinicians certified for the task and manual, but standardized, reading center gradings in DR only reached consistency is a hard task which affect the accuracy and the final result of the study.

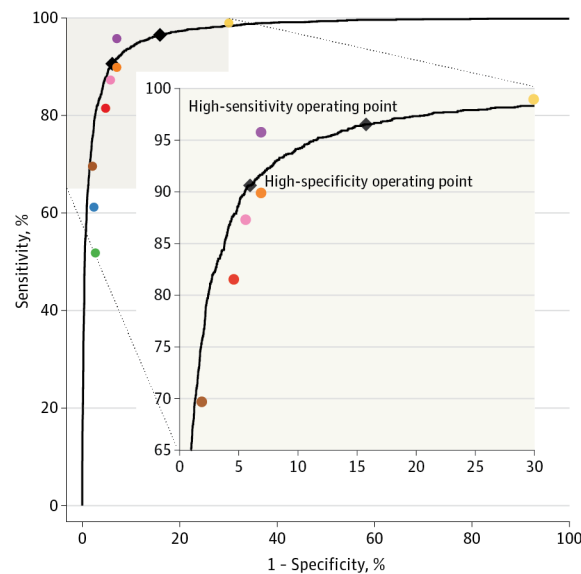


Figure 6: Performance of the Algorithm (Black) and Ophthalmologists (Dots) on the EyePACS Dataset

These results demonstrate that deep neural networks can be trained, using large data sets and without having to specify lesion-based features, to identify diabetic retinopathy or diabetic macular edema in retinal fundus images with high sensitivity and high specificity. This automated system for the detection of diabetic retinopathy offers several advantages, including consistency of interpretation (because a machine will make the same prediction on a specific image every time), high sensitivity and specificity, and near instantaneous reporting of results. In addition, because an algorithm can have multiple operating points, its sensitivity and specificity can be tuned to match requirements for specific clinical settings, such as high sensitivity for a screening setting.

4.1 Experiments

Replicate a scientific paper by following the described methods and algorithms is a cornerstone of science. Replicability is essential for the development of medical technologies based on published results specially when the databases used are publicly available. However, there is an emerging concern that many studies are not replicable, raised for machine learning algorithms. Replication is repeating the method as described in the paper and reproducing as obtaining the result as reported. To balance the level of detail and readability of a manuscript, a replication attempt must follow the standard procedures of the field when details about the method are missing.

In this Survey, we make an assessment on the replicability of a deep learning method used in the domain of detecting diabetic retinopathy in retinal images. We have chosen to attempt to replicate and reproduce the method from Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs by Gulshan et al. [Gulshan et al., 2016], this article had been cited 350 time. We chose to replicate this study because it is a well-known and high-impact study within the medical field, the source code has not been published.

The original paper implements an algorithm for detection of referable diabetic retinopathy in retinal fundus photographs. The algorithm is trained and tested on 118,419 fundus images retrieved from EyePACS data set. The original algorithm's performance was evaluated on 2 test sets, and achieved 2 (AUC) for detecting DR of 0.99 for both the EyePACS and the Messidor test sets. The operating point for high specificity had 90.3% and 87.0% sensitivity and 98.1% and 98.5% specificity for the EyePACS and Messidor test sets, whereas the operating point for high sensitivity had 97.5% and 96.1% sensitivity and 93.4% and 93.9% specificity, respectively.

To replicate the same method used to develop the original algorithm, we used images from a publicly available EyePACS data set for training and validation, and we used a subset from the EyePACS data set and images from the public Messidor data set for performance evaluation. We had to find ourselves the validation hyper-parameters and the image normalization method, as they were not described in the original study. Our objective is to compare the performance of the original detection algorithm to our result after replication, taking into account potential deviations in the data sets, and potential differences in normalization methods and other hyper-parameter settings.

We were not able to replicate the original study, because of insufficient level of detail in the method description in the original paper. Our best effort of replication resulted in a model that was not able to reproduce the results of the original study. Our model's AUC for detecting DR for our EyePACS and Messidor test sets were 0.94 and 0.80, respectively. The operating point for high specificity had 83.4% and 67.9% sensitivity and 90.1% and 76.4% specificity for our EyePACS and Messidor test sets, and the operating point for high sensitivity had 89.9% and 73.7% sensitivity and 83.8% and 69.7% specificity.

The results can differ for three reasons. First, the old published list of hyper-parameters used in the original study [Gulshan et al., 2016] lack important details regarding the validation procedure and the normalization method used, so the original algorithm may have been tuned better. Second, there might be errors in the original study or methodology. The last reason is that we may have missed something when replicating the method by having misinterpreted the methodology. We do not know for sure which of the three reasons has led to our considerably worse performance.

We do not believe our results invalidate the original findings of the study. However, our result gives a general insight into the challenges of replicating and reproducing studies that do not use publicly available data or publish there source code online.

4.2 Effect of ethnicity and race

The density of the background in retinal pigmentation is different for different races. Some fund images are lightly pigmented often called "blonde" fundus and some are heavily pigmented called "slate gray: appearing fundus, both still within the range of normality. Although the lesions are same across all ethnicity, the background color may make them simple or difficult to provide a diagnosis. Many of the DR algorithms have used Kaggle, Messidor and Eyepacs images for training them. These datasets are not representative of different races, thus the performance of the algorithm may differ across different races depending on pigmentation.

Ting et al. [Ting et al., 2017] reported the development and performance of their algorithm in multi-ethnic population and with different fundus camera setups. The algorithm showed consistent results in darker fundus pigmentation retinal images in African American and Indian individuals to lighter fundus in white individuals.

4.3 Open Challenges and limitations

Almost all screening systems have been developed and tested using 2-dimensional images and lack stereoscopic aspects, makes identification of elevated lesion like retinal traction and vitreous traction challenging. Using the information from 3-dimensional imaging in future algorithms may potentially address this challenge.

The training and hyper parameter tuning datasets are often subjected to many variables such as field of view, magnification and image quality. The most important aspect for accuracy and efficiency of an algorithm is the quality of labeled data. The more accurate and robust the labeled data is, the better and more universal would be the algorithm. More evidence on methods of getting high quality label data are required.

Also, the questions on privacy protection are particularly sensitive in retinal images as anonymization is not completely achievable owing to the individual nature of the retinal vasculature which provides a fingerprint-like individual feature. Thus, it is not possible to completely anonymize any medical images.

Almost all current systems for DR detection are based on static data. These algorithms cannot handle the time factor, such as the disease incidence and progression which is really important for patience with diabetes.

The other challenge, is the black-box nature of any DL algorithm. It is challenging understanding how exactly a neural network reaches a particular decision, or to identify which exact features it utilizes. How can the results of a model be properly understood by clinicians and researchers? How can we ensure the reliability of algorithms, if we cannot understand how they operate?

Although ConvNet models have achieved high accuracy in diagnosis of DR, the current challenge is the clinical validation and real time deployment of these models in clinical practice. Most of the studies used training sets from homogenous population of a specific region or a publicly available dataset. Diversifying the dataset, in terms of ethnicities, and camera setups to capture the images will address this challenge to some extents.

4.4 Diabetic Retinopathy screening and its future

Advances in mobile hardware for Deep Learning have enabled iPhone and Android smartphones to run diabetic retinopathy algorithm offline for medical imaging diagnostic support.

Recent novel DL architecture applied to three-dimensional optical coherence tomography (OCT) scans has shown makes excellent appropriate referral recommendation [Mandal, 2018]. DL analysis of OCT scans for morphological variations in the scan, detection of intraretinal fluid or subretinal fluid, neovascularisation have started gaining great interest recently [Schlegl et al., 2018].

The AI devices provides a screening decision without requiring an ophthalmologist to interpret the retinal images, hence can be used by physicians who may not normally be involved in eye care. The integration of AI into healthcare, would be expected to radically change clinical practice with more people getting screened for retinopathy.

The ultimate goal of the future to save healthcare resources and to incorporate AI into a retinal imaging device fundus cameras that can be used in various locations such as pharmacy, optical shop, etc to screen all people with diabetes.

5 Conclusion

Computational modeling for medical image analysis has great impacts on both clinical applications and scientific researches. Recent progresses in deep learning to have shed new light on medical image analysis by allowing discovering morphological and/or textural patterns in images solely from data.

Based on the observations from this study, the development of similar high-performing algorithms for medical imaging using deep learning has 2 prerequisites. First, there must be collection of a large developmental set with tens of thousands of abnormal cases. While performance on the tuning set saturated at 60 000 images, additional gains might be achieved by increasing the diversity of training data (ie, data from new clinics). Second, data sets used to measure final performance (tuning and the clinical validation data sets) should have multiple grades per image. This provides a more reliable measure of a model's final predictive ability.

As deep learning methods have achieved the state-of-the-art performance over different medical applications, its use for further improvement can be the major step in the medical computing field. However, there are still rooms for improvements. First, lessened in computer vision, where breakthrough improvements were achieved by exploiting large amounts of training data, e.g., more than 1 million annotated images in ImageNet [Russakovsky et al., 2015], it would be one direction to build such big publicly available dataset of medical images, by which deep models can find more generalized features in medical images, thus allowing making a leap in performance. Second, while the data-driven feature representations, especially in an unsupervised manner, helped enhance accuracy, it is also desirable to devise a new methodological architecture, with which it becomes possible to reflect or involve the domain-specific knowledge. Third, it is also necessary to develop algorithmic techniques to efficiently handle images acquired with different scanning protocols, by which there is no need to train modality-specific deep models. Last but not least, when applying deep learning to investigate the underlying patterns in images such as FMRI, due to the black-box like characteristics of deep models, it still remains challenging understanding and interpret the learned models intuitively.

References

- [Abadi et al., 2016] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.
- [Abràmoff et al., 2010] Abràmoff, M. D., Garvin, M. K., and Sonka, M. (2010). Retinal imaging and image analysis. *169208IEEE reviews in biomedical engineering*, 3.
- [Abramoff et al., 2016] Abramoff, M. D., Lou, Y., Erginay, A., Clarida, W., Amelon, R., Folk, J. C., and Niemeijer, M. (2016). Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learningdeep learning detection of diabetic retinopathy. *52005206Investigative ophthalmology & visual science*, 57(13).
- [Arunkumar and Karthigaikumar, 2017] Arunkumar, R. and Karthigaikumar, P. (2017). Multi-retinal disease classification by reduced deep learning features. *329334Neural Computing and Applications*, 28(2).
- [Bastien et al., 2012] Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., Warde-Farley, D., and Bengio, Y. (2012). Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590*.
- [Bengio et al., 2007] Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *2007153160Advances in neural information processing systems*, pages 153–160.
- [Bhavsar and Bhavsar, 2016] Bhavsar, A. R. and Bhavsar, A. R. (2016). Diabetic retinopathy.
- [Burlina et al., 2016] Burlina, P., Freund, D. E., Joshi, N., Wolfson, Y., and Bressler, N. M. (2016). Detection of age-related macular degeneration via deep learning. In *IEEE*, pages 184–188.
- [Colas et al., 2016] Colas, E., Besse, A., Orgogozo, A., Schmauch, B., Meric, N., and Besse, E. (2016). Deep learning approach for diabetic retinopathy screening. *Acta Ophthalmologica*, 94(S256).
- [Collobert et al., 2011] Collobert, R., Kavukcuoglu, K., and Farabet, C. (2011). Torch7: A matlab-like environment for machine learning. In *2011BigLearn, NIPS Workshop*, page 10. Granada.
- [Crawford et al., 2009] Crawford, T. N., Alfaro, I. I. I., Kerrison, J. B., Jablon, E. P., et al. (2009). Diabetic retinopathy and angiogenesis. *813Current diabetes reviews*, 5(1).
- [Cuadros and Bresnick, 2009] Cuadros, J. and Bresnick, G. (2009). EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *509516Journal of diabetes science and technology*, 3(3).
- [Decencière et al., 2014] Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., and Erginay, A. (2014). Feedback on a publicly distributed image database: the Messidor database. *Image Analysis & Stereology*, 33(3):231–234.
- [Desbiens et al., 2018] Desbiens, J., Gupta, S., Stevenson, J., Alderman, A., Trivedi, A., and Buehler, P. (2018). Deep Annotated Learning, Harmonic Descriptors and Automated Diabetic Retinopathy Detection.
- [EyePACS, 2015] EyePACS (2015). Kaggle Diabetic Retinopathy Detection competition.
- [Graham and van der Maaten, 2017] Graham, B. and van der Maaten, L. (2017). Submanifold Sparse Convolutional Networks. *arXiv preprint arXiv:1706.01307*.
- [Gulshan et al., 2016] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., and Cuadros, J. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *[object Object]*, 316(22).
- [Hinton et al., 2006] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *15271554Neural computation*, 18(7).
- [Hinton and Salakhutdinov, 2006] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *504507science*, 313(5786).

- [Jia et al., 2014] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *2014 675678 Proceedings of the 22nd ACM international conference on Multimedia*. ACM.
- [Krause et al., 2018] Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G. S., Peng, L., and Webster, D. R. (2018). Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*.
- [LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *436444 Nature*, 521(7553).
- [Lee et al., 2015] Lee, R., Wong, T. Y., and Sabanayagam, C. (2015). Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *17 Eye and vision*, 2(1).
- [Mandal, 2018] Mandal, A. (2018). Google's DeepMind AI could soon be diagnosing eye conditions.
- [Manivannan et al., 2017] Manivannan, S., Cobb, C., Burgess, S., and Trucco, E. (2017). Subcategory Classifiers for Multiple-Instance Learning and Its Application to Retinal Nerve Fiber Layer Visibility Classification. *11401150 IEEE transactions on medical imaging*, 36(5).
- [Melendez et al., 2015] Melendez, J., van Ginneken, B., Maduskar, P., Philipsen, R. H., Reither, K., Breuninger, M., Adetifa, I. M., Maane, R., Ayles, H., and Sánchez, C. I. (2015). A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays. *179192 IEEE transactions on medical imaging*, 34(1).
- [Mohamed et al., 2007] Mohamed, Q., Gillies, M. C., and Wong, T. Y. (2007). Management of diabetic retinopathy: a systematic review. *902916 Jama*, 298(8).
- [Prentašić and Lončarić, 2016] Prentašić, P. and Lončarić, S. (2016). Detection of exudates in fundus photographs using deep neural networks and anatomical landmark detection fusion. *[object Object]*, 137.
- [Quelleg et al., 2017] Quelleg, G., Cazuguel, G., Cochener, B., and Lamard, M. (2017). Multiple-instance learning for medical image and video analysis. *IEEE reviews in biomedical engineering*.
- [Quelleg et al., 2016] Quelleg, G., Lamard, M., Cozic, M., Coatrieux, G., and Cazuguel, G. (2016). Multiple-instance learning for anomaly detection in digital mammography. *16041614 IEEE transactions on medical imaging*, 35(7).
- [Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.
- [Schlegl et al., 2018] Schlegl, T., Waldstein, S. M., Bogunovic, H., Endstrasser, F., Sadeghipour, A., Philip, A.-M., Podkowinski, D., Gerendas, B. S., Langs, G., and Schmidt-Erfurth, U. (2018). Fully automated detection and quantification of macular fluid in OCT using deep learning. *Ophthalmology*, 125(4):549–558.
- [Solanki et al., 2015] Solanki, K., Ramachandra, C., Bhat, S., Bhaskaranand, M., Nittala, M. G., and Sadda, S. R. (2015). EyeArt: automated, high-throughput, image analysis for diabetic retinopathy screening. *Investigative Ophthalmology & Visual Science*, 56(7):1429–1429.
- [Suriyal et al., 2018] Suriyal, S., Druzgalski, C., and Gautam, K. (2018). Mobile assisted diabetic retinopathy detection using deep neural network. In *2018 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)*, pages 1–4.
- [Ting et al., 2017] Ting, D., Cheung, C., Lim, G., and al et (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318(22):2211–2223.
- [UK et al., 1998] UK, P. D. S. U. G. et al. (1998). Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *837853 The lancet*, 352(9131).

- [van Grinsven et al., 2016] van Grinsven, M. J., van Ginneken, B., Hoyng, C. B., Theelen, T., and Sánchez, C. I. (2016). Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *12731284IEEE transactions on medical imaging*, 35(5).
- [Wild et al., 2004] Wild, S. H., Roglic, G., Green, A., Sicree, R., and King, H. (2004). Global prevalence of diabetes: estimates for the year 2000 and projections for 2030: response to Rathman and Giani. *25692569Diabetes care*, 27(10).
- [Winder et al., 2009] Winder, R. J., Morrow, P. J., McRitchie, I. N., Bailie, J. R., and Hart, P. M. (2009). Algorithms for digital image processing in diabetic retinopathy. *608622Computerized Medical Imaging and Graphics*, 33(8).

Ontology based Intelligent Tutoring Model For Analysis and Generation of Arabic Verbs and Derived Nouns

Laila Ibrahim¹, Mostafa Ezzat¹, Mervat Ghith¹, and Magdy Aboul Ela²

Abstract: Intelligent Tutoring Systems are adaptive computer-based systems. They are designed to aid students in learning a specific subject. Learning Arabic morphology is one of the most challenging skills taught to Arabic language learners. The field of researches in intelligent tutoring systems in this area is still limited and did not reach the similar level of other languages due to complexity of Arabic morphology; it is an opportunity to focus on Arabic morphology intelligent tutoring systems. To address this challenge, this paper presents a design and implementation of ontology based intelligent tutoring model for analysis and generation of Arabic verbs and derived nouns to simulate the behavior of the actual educational process. The result of the experimental evaluation demonstrates the feasibility of applying this model. The proposed model has achieved promising results and we recommend using and enhancing this system to get better understanding for Arabic morphology topics.

Keywords: Intelligent Tutoring Systems, ontology, knowledge representation, rule-based Systems, Arabic Morphology.

1- Introduction

Intelligent Tutoring Systems (ITSs) aim at tutoring students in the absence of or besides human tutoring. ITSs provide adaptable support to each student based on modeling of the student knowledge and characteristics [Chakraborty,2010]. A typical ITS generate and/or accepts problems to a given domain and attempts to analyze the student error and build a short-term student model to provide relevant feedback and help. A more advanced ITS builds a long term student model to provide the student with personalized problems according to the assessed level of knowledge.

¹Department of Computer and Information Sciences, Institute of Statistical studies and Research, Cairo University, Giza, Egypt. ²Sadat Academy, Cairo, Egypt.

alzakiadin@gmail.com, most_ezzat@hotmail.com, mervat_ghith@yahoo.com, maboulela@gmail.com

ITSs are characterized by adapting their response to each student according to his needs and characteristics. ITSs involve the student in an active learning environment, aiming at matching the effective learning outcome of an individual tutor [[Souidi, et al,2007]]. These systems contain a large amount of educational knowledge and employ pedagogical methodologies.

Traditional ITSs consist of the following four modules: (a) the domain module, which contains all the knowledge (educational content), (b) the student model, (c) the pedagogical module, which contains information relevant to the various pedagogical decisions and (d) the user interface which enables communication between the user and the system [Woolf,2009].

Ontologies started in the Artificial Intelligence (AI) and Knowledge Engineering field but are being adopted in ITSs. Ontology engineering has been paid more attention in the domain of intelligent tutoring systems (ITS) [Mizoguchi R., et al.,1996] [Murray, T.,1998].

Ontology is an explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them. Many advanced architectures have adopted ontologies for their knowledge representation because it enables accumulation of knowledge and supports knowledge reuse and sharing among different systems [Gruber, T.,1993]. Building educational materials for intelligent tutoring systems is difficult and time-consuming. Thus, it worth building an ITS that can use preexisting ontologies without much effort.

Arabic is a Semitic language spoken by more than 337 million people as a native language. Arabic is a highly structured and derivational language where morphology plays a very important role. Over the last few years, Arabic ITS's has gained increasing improvements, and several state of the art systems have been developed for a wide range of applications such LTS [Shaalán K.,2005].

The main contribution of this work is the development of ontology based model for the analysis and generation of Arabic verbs and derived nouns ITS. In most of the proposed models and systems, teaching domains representations are reduced to pedagogical resources descriptions and the semantic contained essentially in the links between the domain components; very useful for some learning activities, are not

modeled, we realized the effect of teaching domains modeling on the learners' knowledge evaluation and on his progression. The existing models and systems designed for teaching Arabic morphology lack some "intelligence" and don't cover most topics. Indeed it's a promising opportunity to focus on teaching the analysis and generation of Arabic verbs and derived nouns.

In most models and systems, teaching domains representations are reduced to pedagogical resources descriptions and the semantic contained essentially in the links between the domain components; we realized the effect of teaching domains modeling on the learners' knowledge evaluation and on his progress.

Here, we provide a framework for development of knowledge based intelligent tutoring system [Chakraborty S, et al.,2010]. The system accepts domain ontology [Murray T.,1998] and tries to tutor knowledge to the learners and evaluates them according to their answers to automatically generated questions. It also infers the curriculum dynamically based on user's responses and from preexisting relations and hierarchies between concepts in the domain ontology.

The rest of the paper is structured as follows. Section 2 explores the related work that focus on the Arabic tutoring systems. In Section 3, we elaborate on the proposed model approaches based on ontology technology, we outline the ontology that represents the domain model, student model, learning styles, pedagogical model, and interface model. In section 4, we discuss an implementation of a case study for our proposed model. In section 5, we discuss evaluation of our empirical study. Finally, in Section 6, we summarize our conclusions and propose our future work.

2- Related Work

In this section, we discuss related work focused on intelligent tutoring systems specially, Arabic language intelligent tutoring systems.

Despite recent interest in the Arabic language, few ILTSs have been developed for teaching Arabic language [Shaalana K,et al.,2017]. We focus in this section on ITS of Arabic morphology because these systems and models are most related literature help us to the point and understand this domain of research. The most referenced

systems are works done by Shaalan et al., Hafez et al.[Shaalana K, 2005][Shaalana K,et al.,2014], They are available for research and evaluation and well documented.

In this section, we discuss these works to clarify their working method and approach in Arabic morphology ITS.

[Gheith, et al.,1996] Developed Instructional Software for Teaching the Arabic Language (ISTAL) for grade one prep school. The system presents the curriculum as a simple concept associated with a set of generated sentences. Then, the system generated an exercise for the student and the student's answer is automatically evaluated by comparing it to the system's solution.

In [AbouEla, M, 1994] an expert systems-based system was proposed that could be utilized as a tool in constructing the rules of Arabic grammar which in turn could help in building Arabic NLP systems.

[Shaalan K, 2005] developed an ICALL system for learning Arabic based on NLP. The system provided feedback to the learner based on syntax analysis of the learner error and proposed a mechanism of correction by the learner.

In [Khodeir N, 2015] the authors employed a cognitive methodology supporting two modes of operation, active and passive modes. In the active mode, words are generated by ART and the student is prompted to provide the correct roots. AR integrates several techniques for enhanced tutoring. It provides a positive feedback for a correct answer and a negative one otherwise. In the latter case, Prompting Answer Strategy (PAS) is employed, where the student is guided to detect the error by self-explanation.

In [Mahmoud N, 2018] the global structure of ITS consists of mainly four modules: a pedagogic module, a question selector module, an expert module and a student module in addition to a user interface module. But in this system they didn't add expert module because they used Constraints Based Model (CBM) technique. CBM does not teach the learner to employ good strategies but only to keep out of incorrect states.

Unfortunately, almost work has been done in the field of intelligent tutoring systems for generating Arabic verbs and derived noun isn't sufficient enough to satisfy different levels of students of Arabic morphology, they lack providing suitable hints

whenever it's required, they don't diagnose students' errors to identify the cause of the error, they don't cover all topics of derived nouns and infection of verbs.

Developing new model for Analysis and Generation of Arabic Verbs and Derived Nouns surely will improve the perspective of students towards Arabic language, so the proposed model can adapt itself according to the answers of the student, and to be able to generate suitable reaction for each answer either true or false.

3- Proposed System Model

Our objective is to develop an ontology based tutoring model for analysis and generation of Arabic verbs and derived nouns.

In this section, we are going to give a brief description of the basic components of our proposed approach and some of its' basic characteristics of the model.

[Woolf, 2009] has identified five major components: the student model, the pedagogical module, the domain knowledge module and the communication module, and pedagogical module.

Fig. 1 provides a view of the interaction between modules.

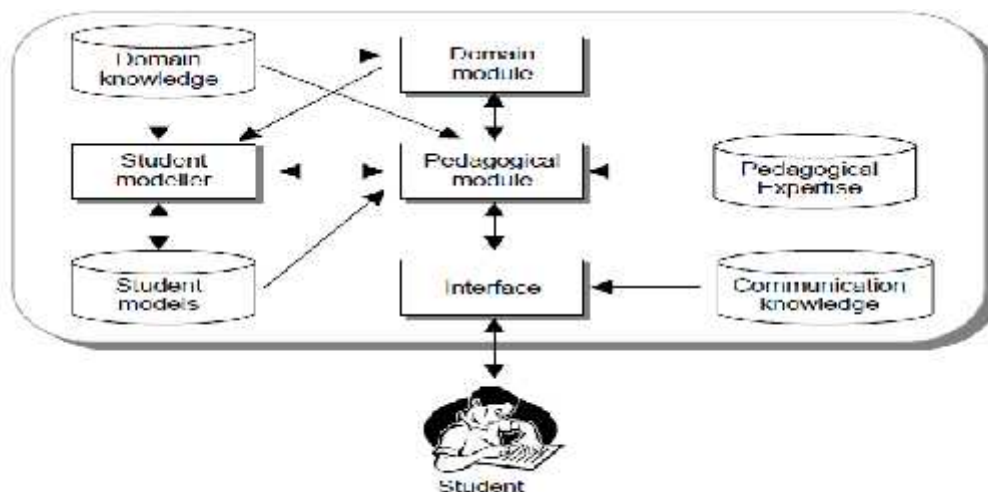


Fig1 Framework Architecture

3-1 The Student Module

It forms a framework for identifying a student's current state of understanding of the subject domain. The knowledge that describes the student's current state of mind is stored in a student model. In order to make any learning environment adaptable to individual learners, it is essential to implement a student model within the system. The student module should permit the system to store relevant knowledge about the student and to use this stored knowledge to adapt the instructional content of the system to the student's needs;

Student model in our model consists of two types of data stored about the student:

- a- **Static data:** Static data includes data about student such as, name, id, e-mail, birth date, and password.
- b- **Dynamic data:** Dynamic data includes the progress of the student throughout the lessons, comments about his last achievements.

3-2 Ontology as a Domain Model

Ontology formalism is usually used to describe the learning material or the learning objects [Zhu, et al.,2018]. The learning object is anything digital that can be delivered across the network on demand such as text, images.

Domain knowledge in an ITS is the knowledge about the subject that's intended to be taught. The domain model organizes the course structure, its various components and the relationship among the components. This model mainly deals with the what-to-teach part of an ITS.[Woolf., 2009]

Our domain subject is organized in three chapters, each chapter contains some lessons, totally 25 lessons, and there are some exercises for each lesson, and an exam at the end of each chapter, then a final exam.

3-3 The Expert Module

The expert model uses the domain knowledge to advise other parts of the system. It may indicate the relative difficulty of curriculum sections or problems, such that the pedagogical module can select the next task. Furthermore, this module aims to provide expert like solutions to problems in the domain to be taught.

3-4 The Interface Module

It controls interactions with the learner, including the dialogue and the screen layouts. This model represents the Graphical User Interface that is responsible for the interaction between the student and the system.

3-5 The Pedagogical Module

The Pedagogical Model specifies the business rules of the system. Each strategy is a rule. This model provides information for the Adaptation Model. Any Pedagogical Model can be thought as composed by phases and each phase has its own strategy like Figure 2 shows.

It contains the knowledge of how to teach, that is, a teaching or tutoring strategy. It orchestrates the whole tutoring process and deals with issues like which topic to present, when to present a new topic, when to present a problem, when to review, and when to offer remedial help.

This module uses information from student model to determine what aspects of the domain knowledge should be presented to the learner and identify the needs for each student. This component provides a model of the teaching process. For example, information about when to review, when to present a new topic, and which topic to present is controlled by the pedagogical module.

The following flowchart declares the role of the pedagogical module.

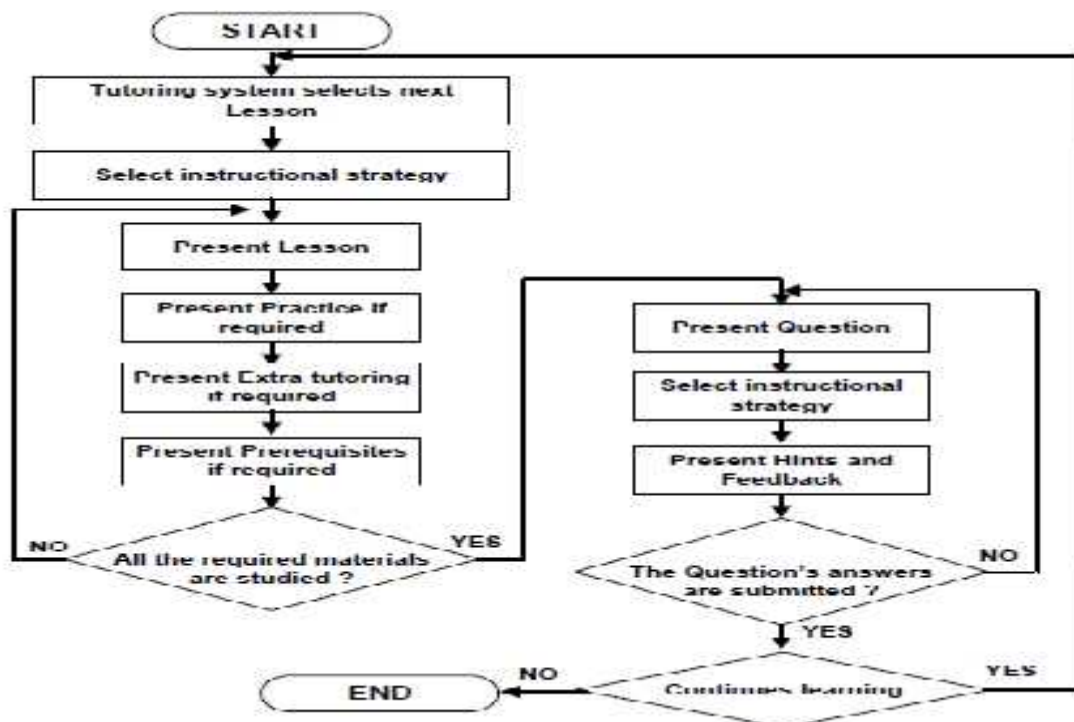


Fig 2 Flowchart of pedagogical module

The following flowchart depicts the whole structure of the proposed model, starting from the entrance of the student to the evaluation stage.

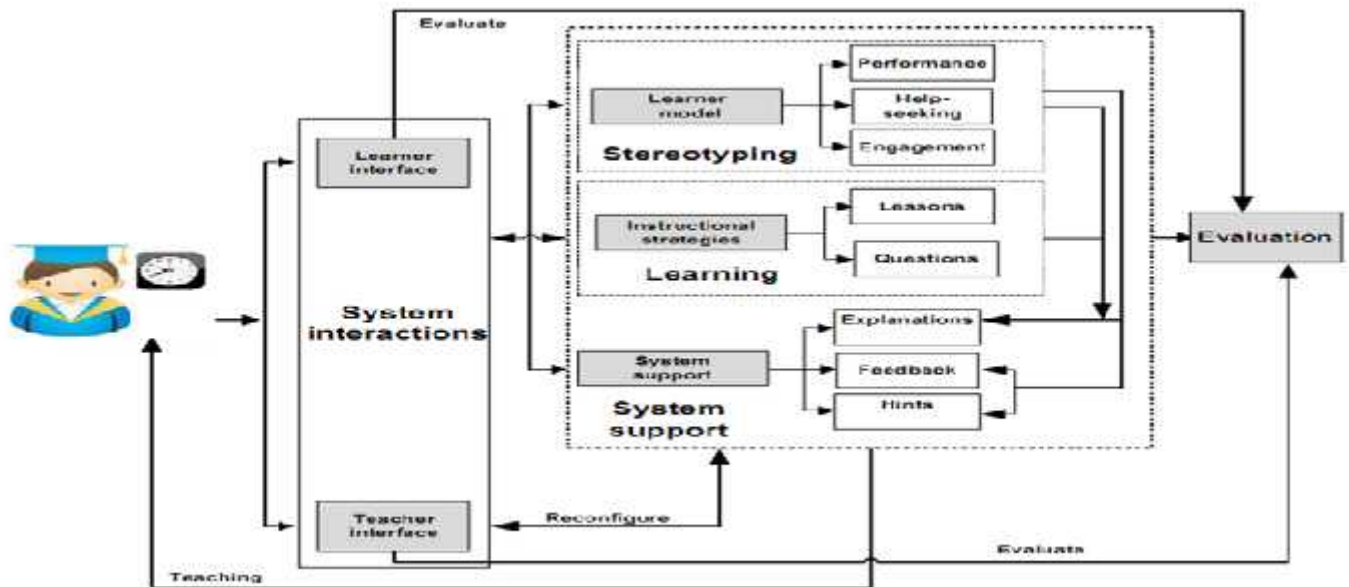


Fig 3 Flowchart of the proposed model

4- System Implementation

4.1 Building the ontology

Ontology is structure of the concepts of that describe the knowledge base.

It contains the concepts of the relationship of knowledge structure as a hierarchical data structure (tree) to describe the components of the knowledge domain.

The proposed ontology has been enriched with a set of rules to enable inference mechanisms (i.e. reasoner) to automatically generate verb inflections and derive nouns.

a- Ontology Classes

In ontology classes stage, we defined 4 main sub classes descending from the main class (الفعل) for Arabic verbs domain based on the main structure of Arabic morphology science as shown in Fig.4.1. Then, we defined the relationship between all classes in order to build class hierarchy.

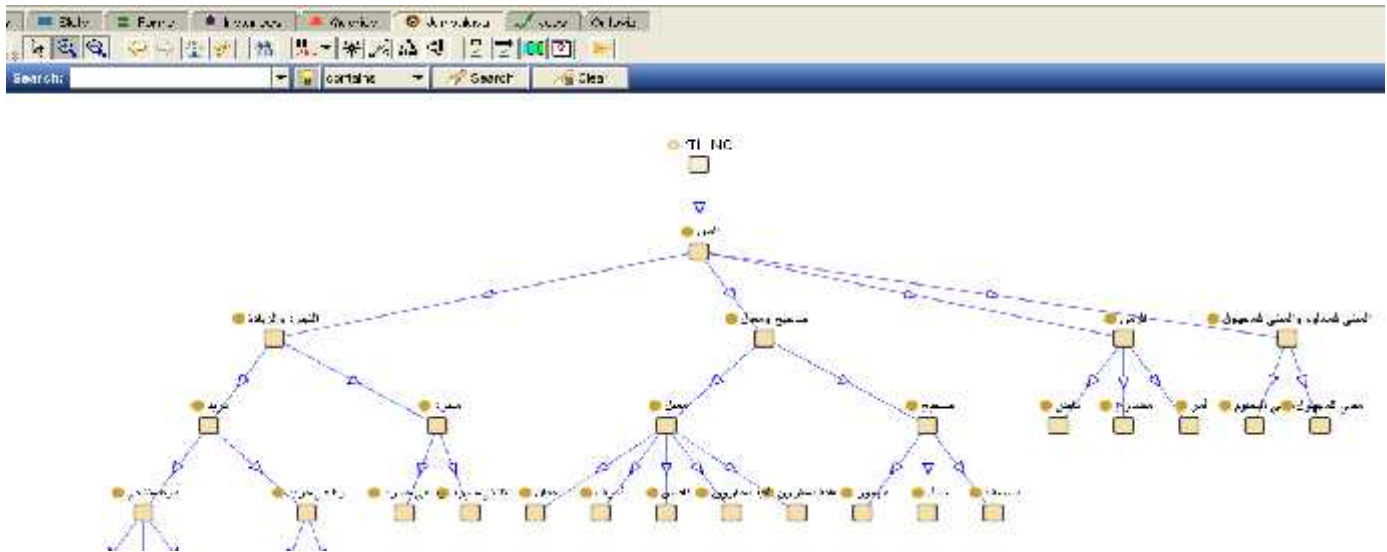


Fig 4.1 Ontology classes of Arabic morphology domain using Jambalaya graphical user interface

b- Classes Properties

Each class has object properties and data types in order to build the relationship between classes and prepare the ontology structure for Arabic morphology.

Referring to Arabic teachers, we can record 20 objects properties related to Arabic verbs domain. Arabic Verb has four main categories:

"الزمن – التجريد والزيادة – البناء للمعلوم والمجهول- صحيح او معتل"

All these classes and sub classes are built in Protégé3.4 and generate frames file that will be ready for JESS tab inferring rules queries to answer questions list in the pedagogical model.

Fig 4.2 Ontology verb taxonomy classes of Arabic morphology domain

c- -Instance creation

In this step, after defining all ontology components and become ready for filling the ontology with an instance. Instance creation allows linking all classes with object properties and data types, which aid to record all classes information for the communication with the student model via pedagogical model. The instances information created through our research from Arabic morphology reference. In our ontology, we created more than 379 instances to analyze and generate Arabic verb inflection and derived nouns.

4.2 The interface

GUI of the system was designed using Net-beans, java swing tool and is linked with Protégé[protégé, 2017] and JESS[JESS, 2013] via java libraries and Jar files, and we used MySQL as database tool. The Interface Module presents the attributes that are related to the characteristics detected by the Student Module.

5- System Evaluation

5.1 System Evaluation according to ontology cycle.

Ontology is used as a formalism to describe knowledge and information that can be shared. Adding more description about the learning style or the role of each knowledge item allows different adaptation technique to be applied such as adaptive presentation and adaptive navigation.

Ontology evaluation is important to ontology development process [Shaalana K,et al.,2017]. It verifies how the developed ontology meets its purpose, and validates the consistency of shared knowledge. We evaluate the developed ontology during each iteration, and between iterations. On one hand, this is a benefit of following an iterative and evolving prototype process. On the other hand, it allows finding missing concepts, attributes, and relation as early as possible.

To assess the effectiveness of the proposed model, a group of thirty students of preparatory stage was formed. The students were given a one-hour lecture explaining how to use the system. Afterwards, they were given a pre-test exam and the grades

were recorded. The students were then allowed to be trained using the system for two hours. They were then given a post-test exam.

a- Evaluation of system

There are several approaches to evaluate the system. The most popular approach is to apply it in a semantic search to test an accuracy of retrieval. The experimental environment of this research uses 30 preparatory stage students to test the accuracy of the system through applying pre- test and post- test. The results of retrieval depends on the level each student. Therefore the accuracy of retrieval of the proposed Arabic language teachers is compared with the results defined by these students. Paired samples *t*-test was applied to compare the grades of the pre-test exam and the post-test exam of the students. The test was performed using Real Statistics. We obtained a *t* value of 8.03. Using this *t*-value and degrees of freedom of 19, we get a *p*-value smaller than 0.00001. Thus, with confidence of interval 05%, we reject the null hypothesis. We found that there is a significant difference between the grades of the pre-test and the post-test exams of the group.

The results of the two experiments (pre-test and post -test) conducted in our study is found in Table1.

T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean	improvement
Pair 1	Post test %	87.47	30	7.300	1.333	37.2%
	Pre test %	63.77	30	14.246	2.601	

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Post test % & Pre test %	30	0.854	0.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Post test % - Pre test %	23.700	10.968	2.002	19.605	27.795	11.836	29	0.000

Table 1 shows the results of the T-test of the system.

b- Discussion of Results

From the results of the empirical evaluation, it is clear that system had a positive effect on tutoring the students. This proves the effectiveness of this novel technique and suggests it as a promising technique in Arabic morphology language tutoring.

c- Challenges

During the implementation, various challenges have been faced us, that we have to overcome. The main challenges were related to our model, we found that there were challenges related to the Protégé tool. In particular, the tool does not have the flexibility to update the name of classes, instances, or properties, which affected the accuracy of the domain model, it consumed a lot of time to modify the classes to update the names of the main and sub classes. For this case, we updated the classes directly from RDF file and resorted to JESS tab to create and modify derivational rules.

5-2 Comparison with other systems

The proposed system is unique in which it's the first tutoring system that includes most topics of Arabic morphology, in a simple way supporting hints and immediate feedback to the students.

The comparison between the proposed system and the other teaching Arabic morphology and grammar approaches is based on the following criteria:

A. Technique and intelligent component:

What is the technique used in, and is suitable or not?

B. Input data:

Does the approach allow the user to accept or decline the entry of data?

C. limitations of the approach:

Does the approach achieve the required tasks?

D. Feedback type:

What type of feedback, this approach provides?

E. Hints:

Does this approach present a suitable hint?

The following table shows and summaries the comparison between the proposed system and other related Arabic morphology and grammar approaches.

	Arabic ICALL (Shaalán,2005;Farghaly and shaalan 2009	A Multi agents ITS for teaching Arabic Grammar 2018	Cognitive tutor of Arabic Word Root Extraction 2017	Our model 2018
Technique Intelligent component	ICALL NLP	Multi agent goals and plans	artificial word generation	Ontology Rule based technique
Input Text	Manual	Words: nouns, verbs, particles bank of questions	Accepts an input word manually and generates its root	Knowledge based for variety of verb types and data base for irregular verbs
Accuracy	NA			Satisfying results
Limitation		Prototype not complete work		Some irregular rules couldn't be handled
Trained Data		Real academic data course of Arabic grammar	Accepts input word and generates its root	Accepts any verb according to the rule and gets all available derived nouns and verb inflections
Subject domain	Arabic grammar	Arabic Grammar of the fourth grade of the elementary stage in Egypt.	Extracting the root (stemming) of Arabic word)	Inflection of verbs and derived nouns
Feedback	Error feedback	diagnose student's problem and give him the solutions to this problem.	Positive feedback for a correct answer and negative one otherwise	Corrective feedback
Hints	Not provided	provided	Provide instant hint	Supports immediate and on demand hints

Table 2 Comparison between the proposed system and other systems

6- Conclusion and Future Work

In this paper, we developed an ontology-based model for some Arabic morphology topics. The salient point in our approach is dynamic extraction of the curriculum based on preexisting relations in the ontology and the responses from the user. This system is developed and implemented to enhance the skills of the student in the adopted domain. The system briefly explains the intended domain which is the Arabic morphology for the preparatory and secondary student. The system supports the explanation process using multimedia facilities such as images, video and text. The system involves question analysis. The knowledge of the Expert Module is important for presenting the correct answer. The system also is supported by graphical user interface to facilitate the interaction with the students. Finally, the system was implemented on the PC/machine Access for the database design using net beans programming languages, Protégé and JESS software packages and Net beans for GUI.

The most challenging part of our research has been the selection of the characteristics to be included in the ontology since Arabic morphology is a highly derivational language.

Our results show that the proposed model that we have implemented on specific domain teaching, as presented above, can be generalized to become reusable for more topics and different domains that can be applied with the structuring process we have defined.

Further work will entail an evaluation with students using more in depth tutorial scenarios, and enhance the GUI to be more adaptive. We further plan to add more rules in the domain model ontology in order to cover as many lessons as possible.

References

- AbouEla, M. (1994). Knowledge-based techniques in an Arabic syntax analysis environment. Unpublished doctoral thesis, Institute of Statistical Studies and Research, Cairo University, Egypt.
- Chakraborty S, D. Roy, A. Basu. (2010). "Development of Knowledge Based Intelligent Tutoring System", Chapter 5, TMRF e-Book Advanced Knowledge Based Systems: Model, Applications & Research (Eds. Sajja & Akerkar), Vol. 1, 2010, pp 74 – 100.
- Gheith, M., Dawa, I., & Afify, M. (1996). ISTAL: Instructional software for teaching the Arabic language. Proceedings of the First International Conference on Computer and Advanced Technology in Education, Egypt, 41 – 66.

- Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5, pp. 199-220.
- JESS. Rule Engine For The Java. Sandia National Laboratories. Livermore, CA. Available at: <http://www.jessrules.com/>. Accessed on January 29, 2013.
- KhairredineBacha, MounirZrigui.(2012). Designing a Model of Arabic Derivation, for Use in Computer Assisted Teaching. *KEOD* : 352-356
- Khodeir N. (2015). Student Modeling for Arabic Language Skills in Intelligent Language Tutoring Systems”, *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 13, No. 11, November 2015.
- Khodeir N. (2017) A Cognitive Tutor of Arabic Word Root Extraction Using Artificial Word Generation, Scaffolding and Self-Explanation. *International Journal of Emerging Technologies in Learning (IJET)*, [S.l.], v. 12, n. 05, p. pp. 36-49, may.. ISSN 1863-0383.
- Mahmoud, M. H. (2018). A multi agents based intelligent tutoring system for teaching Arabic grammar. *International Journal of Education and Learning Systems*,3.
- Mizoguchi, R., et al. (1996). Knowledge engineering of educational systems for authoring system design-A preliminary results of task ontology design, In *Prof. EuroAIED96*, 329-335.
- Murray T.(1998). Authoring knowledge base tutors: tools for content, instructional strategy, Student Model, and Interface Design, *J. of the Learning Sciences*, 7, 1, 5-64,.
- Research, S. (2017). protégé. [online] *Protege.stanford.edu*. Available at: <http://protege.stanford.edu> [Accessed 1 Apr. 2017].
- Shaalaa K.(2005). "An Intelligent Computer Assisted Language Learning System for Arabic Learners", *Computer Assisted Language Learning:An International Journal*, Taylor & Francis Group Ltd., 18(1 & 2): 81-108, February 2005.
- Shaalaa K, Al Emran M.(2014)."A Survey of Intelligent Language Tutoring Systems", *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Delhi-India, pp. 393-399.
- Shaalana K, Albarghothia A, Khatera F. (2017). Arabic Question Answering Using Ontology.3rd International Conference on Arabic Computational Linguistics, *ACLing 2017*, 5-6 November 2017, Dubai, United Arab Emirates.
- Soudi A., Bosch, A., Neumann, G. (2007). "Arabic Computational Morphology: Knowledge-based and Empirical Methods", *Text and Language Technology*, Vol 38. Springer, New York.
- Woolf, B.P. (2009). Building Intelligent Interactive Tutors. Student Centered Strategies for revolutionizing e-learning. Morgan Kaufmann Publ.
- Zhu, X.-H., Wu, T.-J., & Chen, H.-C. (2018). An Interoperable Model for the Intelligent Content Object Based on a Knowledge Ontology and the SCORM Specification. *Journal of Educational Computing Research*, 56(5), 723–749.

A Survey of Sign Language and Machine Translation Systems

Mohamed Amin¹

Ammar Mohammed²

Hesham Hefny³

Abstract

This paper explores the application of data-driven machine translation (MT) to sign languages (SLs) and shows the different approaches of MT systems that can facilitate the communication between Deaf and hearing people by translating information into the native and preferred language of the individual. The paper introduces SLs, focusing on American Sign Language - the native language of the Deaf in America. The description of their linguistics and mechanics including similarities and differences with spoken languages are also described in the paper. Moreover, the paper summarizes previous approaches of SL MT, and focuses on the pros and cons of each approach. Initial experiments in the novel area of example-based MT for SLs are discussed and an overview of the problems that arise when automatically translating these manual-visual languages is given. Following this, an overview of Sign Language Machine Translation system based on rule-based Approaches (RBMT), Example-Based (EBMT), Statistical Machine Translation (SMT) are also discussed.

Keywords: Sign Language, Rule-based Machine Translation, Example-Based Machine Translation , Statistical Machine Translation.

1. Introduction

This paper is a survey of several of the most successful Text to Sign Language (SL) machine translation systems. This survey will first introduce the special properties of SL that make it a challenging subject of computational linguistic study and will then frame the problem of Text to SL translation. After introduced the four systems in this survey, they will be compared

¹ Department of Computer Science, Institute of Statistical Studies and Researches, Cairo University, Egypt
Mohamed.amin.mfa@cu.edu.eg

² Department of Computer Science, Institute of Statistical Studies and Researches, Cairo University, Egypt
Ammar@cu.edu.eg

³ Department of Computer Science, Institute of Statistical Studies and Researches, Cairo University, Egypt
Hehefny@cu.edu.eg

in a thematic fashion according to several criteria important to a successful Text-to-ASL translation system. Finally, each of the systems will be discussed in terms of a particular strength or interesting approach it has taken to the SL MT task. Several researches have shown that SL has a distinct grammar, vocabulary, and structure. And its visual nature allows it to use linguistic phenomena not seen in any spoken language. Unlike traditional natural languages that use spoken sounds or written symbols, the SL relies on multiple channels of handshapes, movements, facial expressions, and other non-manual signals (NMS) to convey its meaning. SL signers will use the space around them for several (grammatical, discourse, and descriptive purposes). Standard sign languages (SLs) are known as Deaf and Dumb languages. SLs are gestural languages that involve symbolic encoded message for communication without speech channel. They are unique in some ways, because they cannot be written like spoken language. SL differ in vocabulary and grammar from country to country. Even within one country, sign language can vary from region to region like spoken languages. *“Gestures are powerful means of communication among humans. Among different modality of body, hand gesture is the most simple and natural way of communication mode. Real time, vision-based hand gesture recognition is more feasible due to the latest advances in the field of computer vision, image processing and pattern recognition but it has yet, to be fully explored for Human Computer Interaction (HCI). (Archana and Gajanan, 2014)”* The translating between languages has become very important. Most people want to acquire and to disseminate knowledge in their own language and in other languages. In an attempt to find a way to translate more material quicker and easier, researchers have developed computer machine translation systems, which have become tools to aid the translation process. Machine translation (MT) is automatic translation from one language to another by computer. Researchers have been studying the process of machine translation for over 60 years. And the research on the SL is still a relatively new area compared to research on languages. A significant study of SL linguistic structure only started in the 1960s with the founding work of Stokoe (1960), whose research on American Sign Language paved the way for social recognition of SLs as real languages. More recent acknowledgment of this is included in the works of linguist such as Pinker (1994) and Chomsky (2000). Increased recognition of SLs as fully-formed, independent languages following political

acknowledgment, such as the Resolution of the European Parliament in 1988, has led to some level of research being carried out on SLs in most countries. Usually it is the national center for Deaf Studies or other Deaf associations that investigate the sociological, educational, cultural and linguistic aspects of SLs. In Ireland, for example, the researchers at the Centre for Deaf Studies engage in valuable ISL linguistic research, and members of the Irish Deaf Society have produced a book outlining the employment and poverty problems experienced by members of the Dublin Deaf Community. Common misconceptions that sign languages are universal, if they were, there would be no barrier between the deaf communities. There are 121 sign languages for deaf people around the world (The ethnologist World Language) (Achraf, 2017). The majority of these languages are distinct languages in their own right that have evolved Either naturally in the deaf communities themselves or were originally borrowed Other SL. Even within SLs from different countries, there are some variations. American Sign Language (ASL), German Sign Language / Deutsche Gebärdensprache (DGS), Ireland Sign Language (ISL). Another common misconception is that SL are grammatically dependent on spoken languages. if that were true, then the role of MT would be only a direct dictionary look-up system for finding the gesture sequence corresponding to each spoken language word would be needed. Like word-for-sign. There is sort of SL that follow the grammar and often the morphology of the spoken language. An example of this is Signed Exact English (SEE). These versions of sign languages follow to spoken language grammar (in this case, English), where there is a sign for each word of the spoken language sentence and often for each morpheme as well. This form of SL is sometimes used in the media as a means to interpret spoken news to a SL. to reduce the time of translation process ignoring grammar structure, best use of the signing space and simultaneous spatial linguistic phenomena of SL. Because the manual signing process itself is hard and slows down the communication of information. SLs are independent of spoken languages, make the best use of the way of communication and thus bring to light interesting issues for MT to tackle in cross-modal translation. The main goal of this paper is to present a survey on discuss the translation techniques for sign language translation. The rest of this paper is organized as follows: Section 2 presents background on sign languages. Section 3 discusses sign language and machine

translation. Section 4 presents overview of sign language machine translation. And the last section presents the conclusion of this work.

2. Background on Sign Languages

In this section will briefly introduce the linguistics of sign language based on the articulation and structure, manual and non-manual features, signing space and time adverbials, transcription and notation systems.

2.1 Sign Language Linguistics

Now that we have shown that there are multiple SLs and that they are distinct from spoken languages, we will discuss the linguistics of SLs in general and point out the relevance of these issues for Machine Translation.

2.2 The Articulation and Structure

The articulators are differed in SLs compared to spoken languages. The primary articulators in spoken languages are (the throat, nose and mouth) while the main articulators in SLs are the (fingers, hands and arms), as shown in figure (1).

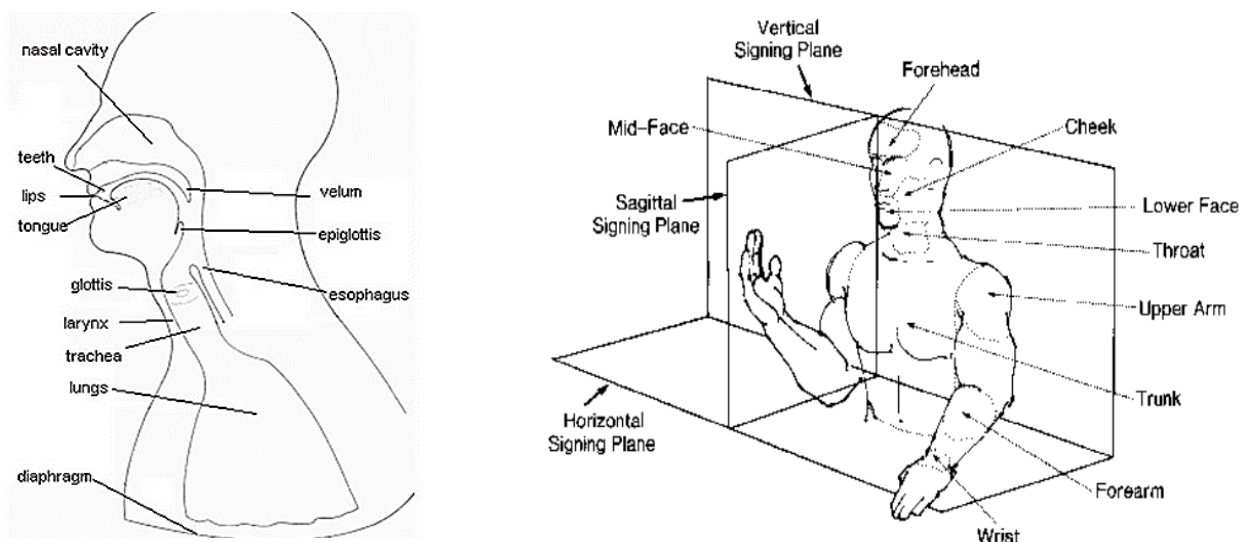


Figure (1) : The Articulation

The signs in SL are similar to morphemes in spoken languages and the articulations of the hands and body can be categorized as phonemes like those in spoken languages (Stokoe, 1960). However, unlike speech, these phonemes are not linear and sequential, but rather occur simultaneously. There are five categories of phoneme in an articulated sign “five parameters in SL”: (Hand shape, Palm Orientation, Location, Movement and Non-manual signals / facial expressions). Many of these articulators in SLs can be thought of as moving through the space in front of the signer: the locations of the hands, the destination point of the eye gaze, the destination point of the head tilt, etc. The Structure Sign language (will take ASL to demonstrate). ASL sentences have three parts (time, topic, comment). The time is when it happened, the topic is what you're talking about and the comment is so why is it important

2.3 Non-Manual Features

Non-manual signal is a parameter of a sign that has meaning but it isn't made with hands. Non-manual signals (NMS) play an important role in specifying tense in SL. The non-manual signals, which include (facial expression, eyebrow movement, movement of the eyes/cheeks, mouth patterns, tilting of the head, movement of the upper body and shoulder movements) Without NMS, a statement will be meaningless regardless of whether the syntax is in the proper order (Achraf, 2017). SL relies on non-manual signals to convey the difference between declarative, imperative and interrogative sentences, as shown in figure (2).

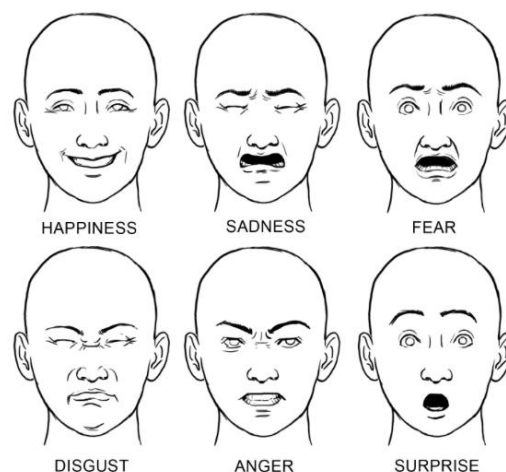


Figure (2) : Non-Manual Facial Expression

2.4 Signing Space and Time Adverbials

An important concept to understand is the ASL sign space. This refers to the area in which most signs are made in normal conversation. Signing outside the sign space is uncomfortable on your hands, wrists, and shoulders. Generally, most signs are centered to the right or left of your middle torso. Sign languages are gestural languages that are articulated in such a way to make the best use of the space in which articulation can take place, in a similar way that oral speech makes best use of the vocal tract, nose and mouth. All manual and non-manual articulations take place within this space. SLs have developed to make the best use of the visual communication channel. ASL give information about time with a specific set of gestural that are produced in relation to what is called a time line (Morrissey and A. Way, 2006). as shown in figure (3).

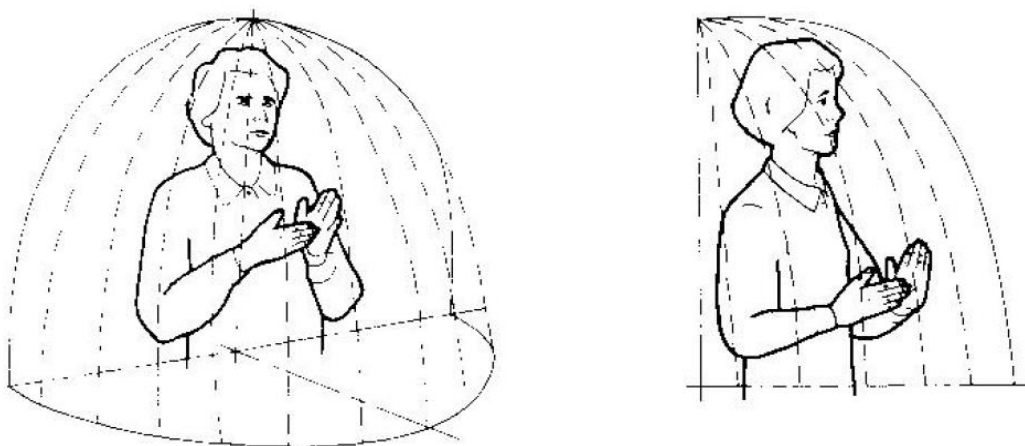


Figure (3) : Diagram of Signing Space

2.5 Transcription and Notation Systems

2.5.1 Stokoe Notation

The Stokoe notation system was developed in 1960 by William Stokoe (Stokoe, 1960) to write American Sign Language (ASL), and later adapted to write the British and Australian sign languages. It was the first script created for phonetic writing of a signed language, even if its primary purpose was not to serve deaf people to note their language, and that it has never been used for this purpose. On the contrary, it has always been devoted to research and linguistic analysis. Stokoe was the first linguist to study and offer a description of the language of the signs in terms of phonemes that he called “cheremes”. He initially describes three factors, namely the tab (tab) referring to the location of a sign, the designator (dez) referring to the shape of the hand, and the signatures (sig) referring to the type of movement to articulate. An example of the Stokoe notation for the sign “DONT-KNOW” is illustrated by the figure (4). This notation uses the Latin alphabet and the numbers to represent the form of the hand. Glyph icons are used to represent the positions of the hands and Movements. The first symbol indicates the place of articulation that is on the forehead. The Letter “B” indicates that the hand takes the form B and the letter T indicates the orientation of the hand, being the palm turned towards the signatory. The letter “x” indicates the presence of a contact. Then, the letter “p ” indicates that the palm is Turned down. The “T inverse” Indicates that the palm of the hand is opposed to the signatory.



Figure (4) : Diagram of Signing Space

This notation is written in a linear way from left to right. The symbols are arranged in a standard order: the location, the shape of the hand, the orientation and the Movement. This differs from the notation “SignWriting” Character-based Which is noted from top to bottom. This notation system is a description graphic where the signs are stored images, making it hard to recognize from the sign by a machine, from the image. Also, the Stokoe notation does not care about Non-manuals gestures such as facial expressions and look.

2.5.2 SignWriting

An alternative method was developed in (Valerie Sutton 1995) which is called “SignWriting” (Stuart, 2011). This approach also describes the phonologically sign language, But, unlike others, it was developed as a visual writing system That allows you to read, write and type any Sign language of the world. This system uses visual symbols to represent the configurations, Movements and facial expressions of all sign languages. These drawings in Simple stroke, as shown in figure (5) make the system easier to learn, because they are more intuitive and visual and then connected to the signs themselves. The system “SignWriting” is now taught to deaf children and Adults from around the world as a handwritten version of sign language. The system is not yet widely used, but its ergonomics and the speed at which it was adopted gave rise to the publication of books for learning the system. The linguistic detail of the sign language is encoded in the drawing line.



Figure (5) : Example of SignWriting: deaf

This graphical nature of SignWriting is actually a practical notation for human, easy to comprehension and reading. However, this nature Graph also makes the notation difficult to analyze with a computer. The notation should be adapted to use only standard ASCII characters or be coded in a markup language such as XML in order to be useful for computer animation.

SignWriting Markup Language Abbreviated as SWML, an XML-based format, has been proposed. This notification has become more exploitable by software such as V-signs that use SWML-encoded resources to generate an animation through an avatar. An example of a notation is shown in figure 6.

```

1  <signbox>
2    <seq>01-05-034-01-02-01</seq>
3    <seq>02-08-012-01-01-02</seq>
4    <seq>04-04-001-01-01-01</seq>
5    <sym width="36" height="35" left="-18" top="-18">04-04-001-01-01-01</sym>
6    <sym width="17" height="20" left="-8" top="19">01-05-034-01-02-01</sym>
7    <sym width="26" height="23" left="-10" top="-43">02-08-012-01-01-02</sym>
8  </signbox>

```

Figure (6) : Example of SignWriting: deaf

Figure (6) shows the SWML file corresponding to the sign “DRINK”, as it is stored in a SW text file generated by the SignWriter editor. Each element of a SW sign is called "symbol". These symbols are encoded by a sequence of 6 integers to find the equivalent of the corresponding character. Although the signed languages are of a higher semantic density than written languages, if a comparison made between the letters and the symbols, between the signs SW and the words, it is here necessary to use a line for each letter, and on average 8 lines for each word. This encoding is therefore very complex from a computer point of view, but also from a point of view of the calculation time and the storage space required. In addition, for computer processing, it makes use of image processing techniques since the SWML file contains no linguistic information, only the locations in space are described.

2.5.3 HamNoSys

Another explicit and graphic notation system for sign language is Hamburg rating system “Hamburg Notation System” or “HamNoSys” (Hanke, 2004). HamNoSys is language independent and applicable to any SL that uses a set of language-independent symbols that represent the phonological characteristics of sign language. It is rooted from the system of Stokoe. And the main purpose of this notational system was to be a writing/transcription system

useful for researchers of signed languages, focused on how to specify the handshape, palm orientation, and movement details significant for performing sign language, it provides written support for researchers on sign languages in order to store them.



Figure (7) : Example of SignWriting: deaf

This example as shown in figure (7), illustrates most of the description categories used in HamNoSys notation. The first colon-like symbol indicates that the hand arrangement is a twohanded sign where the hands are parallel but not mirroring each other. The square brackets denote a grouping of simultaneously occurring movements. Within these braces, the oval symbol with four vertical lines indicates one hand has four fingers extended and aligned together, the same symbol repeated with a fifth diagonal line denotes the same hand shape but with the thumb extended this time. The horizontal line with the dash through it in between these two hand shapes shows the movement is repeated from the starting point of the original movement. Outside of the brackets, the next symbol, a horizontal line with an upward facing arrow above it, indicates that the hand orientation is away from the body, while the next symbol, an oval with the lower half shaded, signifies that the palm of the hand is facing down. The final symbol, a vertical line with a clockwise circular arrow to its right, shows that the movement of the hands is circular, moving vertically and away from the body. HamNoSys is often used for recognition and automatic generation (e.g. SiGML). This process aims to generate three-dimensional animation via conversational agents. From MT perspective, HamNoSys Notation and Stokoe Notation are not suitable notation to use in machine translation process.

2.5.4 SiGML

The “Signing Gesture Markup Language” system says SiGML (Elliott et al., 2004) was initially developed in the ViSiCAST project (Bangham et al., 2000) which seeks to develop a set of tools used for the synthesis of gestures of sign language. SiGML is based on HamNoSys notation, that presents a decomposition of the description of a sign similar to that of HamNoSys. It is designed to accommodate to the general XML standards and represents the notations of HamNoSys. Furthermore, it also supports several notation possibilities for the non-manual signals, the annotation is pretty straight forward, such as the shape of the hand, the orientation, the location, and the movement. This is a category that includes not only location changes, but also shape and orientation. It is also providing not only non-manual signals specifying the behavior of the articulators of the face, such as the mouth, eyes, eyebrows, the gaze, but also non-manual body movements, such as winks, nods and body tilt. As can be seen in figure (8) this SiGML encoding for the sign “I” in ASL represented with both Manual and non-Manual signs.

```

<hamgestural_sign gloss="I">
  <sign_nonmanual>
    <mouthing_tier>
      <mouth_picture picture="I x"/>
    </mouthing_tier>
  </sign_nonmanual>

  <sign_manual>
    <handconfig handshape="finger2" thumbpos="across"/>
    <handconfig extfidir="il"/>
    <handconfig palmor="r"/>
    <location_bodyarm location="chest" contact="touch"/>
  </sign_manual>
</hamgestural_sign>

```

Figure (8) : Example of Notation of the sign "I" in SiGML and HamNoSys

The advantage of this annotation that is easily to understand and readable to both human and computers. But, its disadvantage is the need for an extra stage of translation.

2.5.5 SML

The Sign Modeling Language System (SML) (Jemni and O. El Ghoul, 2008) is a descriptive language XML . SML is used to store descriptions of signs, that use to do simulation with 3D-animation using avatar. This simulation can be used in different systems like robot animation, chatbot, online learning for deaf. The animation, according to SML, is a set of movements or rotations of groups of joints classified according to their movement above the axis of the Times.

2.5.6 Textual Annotation “Gloss”

A System or code like (Braille, fingerspelling, Morse Code). Gloss (Klima, 1979) system used to annotate, representing, describe sequences of signs (a visual gestural language) based on labels (words) written form. Gloss is a way of writing about or representing one language in another language form. It is an interlinear translation., used by linguists for transcription, people who are studying language when an exact way to write down everything that they see signed, it's also used by learners of SLs for analysis. And consider a good way to focus on the grammar and word order separate from vocabulary, so you don't get distracted with how to produce the signs. Gloss is written above the natural words using CAPITAL letters. The table below shows pairs of (English, ASL) sentences.

English Sentences	ASL Gloss
I went to the football game Friday night.	FRIDAY NIGHT FOOTBALL GAME I GO
This is Mary. She likes Chinese food.	IX-Mary fs-MARY. CHINA FOOD IX-Mary LIKES
What is your name?	NAME YOU WHAT ^{WH}
He doesn't like pizza.	PIZZA IX-boy DOESN'T-LIKE
Help me.	HELP-ME (one sign)
See you later.	SEE-YOU-LATER (one sign)
Don't know.	DON'T-KNOW (one sign)
Today is Friday, October 28th.	NOW+DAY FRIDAY fs-OCT 28

There are some signs that are just so powerful that one English word isn't enough to represent them like a directional sign such as “help me” if we wrote in glass “HELP-ME” as two separate words we'd be tempted to sign it that way but “HELP-ME” come straight towards you so it's one

sign so we're going to hyphenate it. Another example “SEE-YOU-LATER” one sign not three separate ones even though we need three words in English to represent that concept. sometimes the opposite is true we need two or three signs to talk about one English word like for instance “Teacher” is two signs “TEACH” + “PERSON” so when we're going to combine those signs together we're going to use a plus in between them. For example, “Today is Friday, October 28th.” today is two signs “NOW+DAY” and then we're going to sign Friday then we fingerspell “fs” October OCT “fs-OCT” and sign the number 28. Gloss provide nice way to describe a Non-manual signal. NMS written above SL gloss separated by horizontal line. For example:-

English Sentences	ASL Gloss
I Don't know.	_____confused_____ ME DON'T-KNOW

to gloss a question the same way, write the English in the gloss and then draw a line above the gloss question like any non-manual signal for a yes/no question we write a “q” above the line, for a “wh” question use “whq” above the line.

_____q_____

For Yes/No questions: YOU

_____whq_____

For Wh- questions: WHERE

English Sentences	ASL Gloss
I love that book, have you read it?	_____love_____ _____q_____ IX-book BOOK ME KISS-FIST, YOU FINISH READ

Sometimes the whole sentence needs the same non-manual signal. Other non-manual signals only apply to part of the sentence or may need more than one non-manual signal in the same sentence that means we only need to put the line over the sign the non-manual signal applies to.

English Sentences	ASL Gloss
I really hated that movie.	_yuck!_ IX-movie MOVIE I HATE
Why are you tired?	____wh-q____ YOU TIRED WHY
Is that your father?	____Q____ IX-man YOUR FATHER
I'm excited today is Friday!	____☺____ NOW+DAY FRIDAY. I EXCITED

3. Sign Language and Machine Translation

Machine translations are form of computational linguistics and language engineering, which uses tools to translate text or speech from one language to another. The two most common engines are rule-based and statistical these engines differ in the way that they process and analyze content, that are often combined within the same system and known as hybrid-MT. and will focus to them in details in next section.

3.1 Machine Translation Approaches

Machine translation (MT) systems have different architectural designs and most of them can be classified into three basic designs: direct, transfer, or interlingua. **Direct approach** (Dictionary-based machine translation) do processing to individual words of the source language string, and the translation is achieved to the target language without performing any form of syntactic analysis on the original input text. In another words it is word by word translation system, without much correlation of meaning between them. Dictionary lookups may be done with or without morphological analysis. **Transfer approach** do analyze the input text to some syntactic or semantic process, and then apply special set of “transfer” rules to read the information of the source language structure and generate a corresponding syntactic or semantic structure in the target language. Afterwards, transfer process does mapping based on source syntactic structure to target syntactic structure form. The analysis of the source language input in this approach is usually shallow comparing with interlingual approach. **Interlingual approach** is an abstract

away from language-specific characteristics, to creates a “language-neutral” representation, can be used as a “pivot” representation in the translation. The two major advantages of Interlingua method are: The more target languages there are, the more valuable an Interlingua becomes. Interlingual representations can also be used for multilingual applications. The Interlingual method do analyze of input text one step further: the source is analyzed and semantically processed to produce a typically language-independent semantic representation structure called an “interlingua”, and then a generation component produces the target-language surface form from this starting point. The figure (9) shows the different machine translation approaches and how processing done from source language to target language.

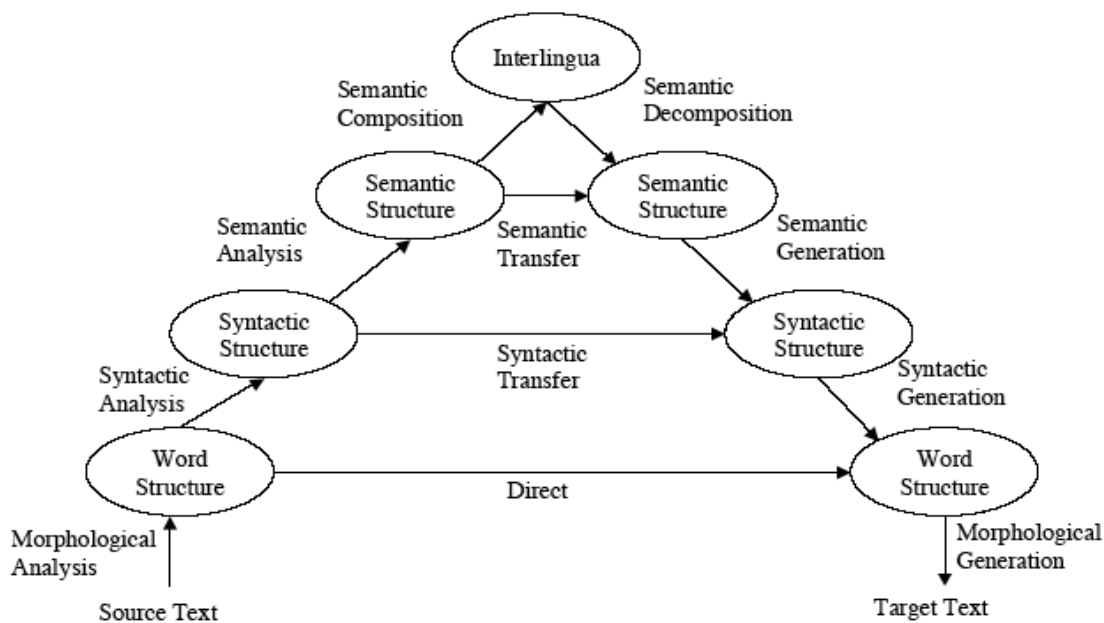


Figure (9) : Machine Translation Architecture

3.2 Machine Translation Types

There are different kinds of machine translation (MT). **Rule-Based (RBMT)** is classical Approach of MT, based on linguistic information about source and target languages. RBMT translation system on the basis of morphological, syntactic, and semantic analysis of both the source and the target languages. RBMT engines use set of linguistic rules and exceptions to break down the content. It produces more predictable output for terminology and grammar through the use of customized terminology lists to fine tune the engine the ability to correct every error with a targeted rule. Rule-based engines don't need a large and structured set of texts also known as a "bilingual corpus" to create the translation system. The direct, transfer and interlingual machine translation methods all belong to RBMT but differ in the depth of analysis of the source language and the extent to which they attempt to reach a language-independent representation of meaning or intent between the source and target languages . This sort of systems is heavily modified content and needs experienced linguists to apply linguistic rules for the translation process in the MT system. **Example-Based (EBMT)** that tries to match (input sentence) to (similar phrases and sentences) in the corpus, retrieving their translation for the output. EBMT is the most suitable approach to process languages whose grammar is barely known, and which has limited resources (in our context, SL). Even with a limited data set, the output is efficient compared to other approaches. All inputs and outputs in an EBMT system are human-readable, especially comparing with the statistical approaches. **Phrase-based (PBMT)** which translates small sequences of words or phrases. **Statistical Machine Translation (SMT)**: SMT used statistical models to generate the translation of the source content statistical engines don't analyze text based on language rules instead the statistical model is built by analyzing bilingual corpus (parallel corpus) requires an appropriate volume of bilingual content to train a model for translating sentences from the source language to the target language and finds the highest probability of any given translation. **Hybrid MT** which is a combination of multiple MT approaches. **Neural Machine Translation (NMT)** : Neural networks are statistical learning models that were first used in speech and image recognition technology its application and machine translation, enables empty engines to train themselves using a process which is similar to the way that the human brain works through trial and error this process is called deep learning

and comes from the principles which have been established through the implementation of big data analytics.

4. Overview of Sign Language Machine Translation

Sign languages do not have a widely accepted written form in which to create a parallel corpus, thus sign languages were largely excluded from MT research until the twenty-first century. Today, many researchers worldwide are investigating sign language MT (SLMT) using specialized textual glosses for representing sign languages (Mary, 2015).

4.1 Rule-based Approaches

The following sections discuss systems using the paradigms described above in an SL context.

4.1.1 ViSiCAST Project

ViSiCAST (Bangham et al., 2000) is a project funded under the European Union Fifth Framework which is part of the Information Society Technologies (IST) program. The system for translating from English text into British Sign Language (Marshall & Sáfár, 2001). The approach of the project used the **CMU Link Parser** for analyzing the input English text, and then uses **Prolog declarative clause grammar rules** to convert this linkage output into a Discourse Representation Structure (DRS). During the first half of the translation process, Head Driven Phrase Structure rules are used to produce a symbolic SL representation script. This script is in the proprietary “Signing Gesture Markup Language” SiGML a symbolic coding scheme for the movements required to perform a natural sign language, as shown in figure (10).

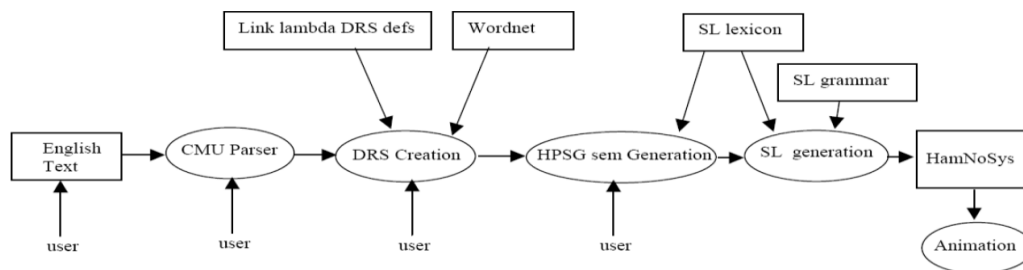


Figure (10) : ViSiCAST Structure

4.1.2 ZARDOZ Project

The ZARDOZ framework, proposed by (Veale et al., 1994) (and somewhat implemented) English-to-Sign-Languages translation framework using a set of hand-coded schemata as an interlingua for a translation component. ZARDOZ framework embrace the knowledge-intensive interlingua approach, that mean try to model the input text directly , and the reasons of that are to separate the processing of source and target languages to allow the ability to generate several different sign languages an interlingual stage of representation allows the system to bring common-sense inference to bear upon the translation process. Some of the research foci of this system were the use of AI knowledge representation, metaphorical reasoning, and a blackboard system architecture. So, the translation design is very knowledge-based and AI-reasoning heavy. During the analysis stage, English text would undergo sophisticated idiomatic concept decomposition before syntactic parsing in order to fill slots of particular concept/event/situation schemata. The advantage of the logical propositions and labeled slots provided by a schemata-architecture was that commonsense and other reasoning components in the system could later easily operate on the semantic information. Because of the amount of hand coding needed to produce new schemata, the system would only be feasible for limited domains. To compensate for these limitations, if a schema did not exist for a particular input text, the authors suggest that the system could instead perform word-for-sign transliteration (producing Signed English output), as shown in figure (11).

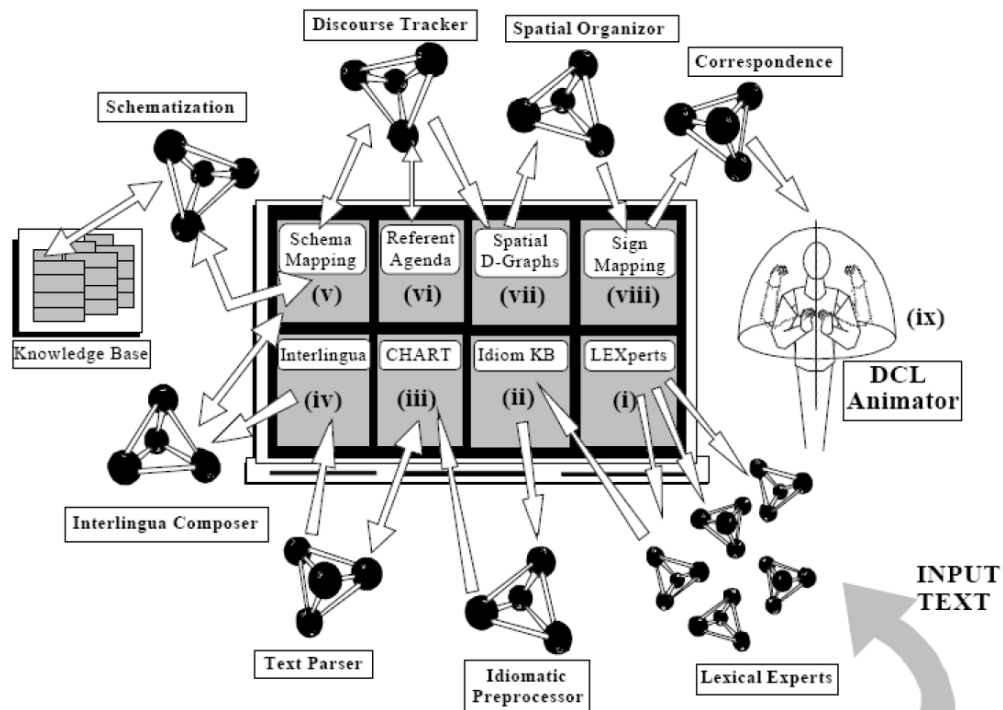


Figure (11) : ZARDOZ Structure

4.1.3 ASL Workbench Project

The ASL Workbench was proposed by (DíArmond Speers, 2001). which is based on a thorough understanding of modern ASL linguistic research. The framework incorporated modern research on ASL phonological models and apply transfer approach using a lexical-functional grammar (LFG) which employs complex hand-engineered rules and requires a tremendous amount of human interaction in another words analysis of the input English text into a functional structure, hand-crafted transfer rules for converting an English f-structure into an ASL one, and LFG rules to produce ASL output. The framework uses a transfer-specific lexicon to map English words/phrases to analogous ASL signs/phrases. The framework interacts with the user of the system and ask for advice if encounters difficulties in lexical choice or other translation tasks. In fact, the system cannot operate correctly without human intervention. Because it designed to be a tool to assist professional translators in converting from English to American Sign Language.

The framework also provides a very simplistic discourse model from the English input (consisting of a flat list of discourse elements and their spatial locations if yet specified), but all reference resolution must be performed manually by the human operator. The ASL Workbench partially implemented English-to-ASL as machine-assisted translation system.

4.1.4 TEAM Project

TEAM project (Zhao et al, 2000) is a framework that translate an English text to ASL. The system built at the University of Pennsylvania. The project was developed using a transfer approach for translation. The input data are analyzed and distinguished in syntactic and morphological information. Subsequently, find an intermediate representation, abbreviated to RI, produced in glossy notation with some built-in parameters, such as facial expressions. During the first phase being the generation of IR and the word order analysis in English, the order of the signs will be detected. Then, the tree grammatical according to the "Synchronous Tree Adjoining Grammar (STAG) " is used to map information from English to ASL. **In the second phase** of the translation process, the IR information is fed into the synthesized sign. In this project, the focus is on modeling Fine movements including aspectual properties and additions in a way, thanks to an approach based on the notation LMA (Laban Movement Analysis).

4.1.5 TESSA Project

Another extension of the VisiCAST (Bangham et al., 2000) project is the project TESSA (Text and Sign Support Assistant) (Cox et al., 2002) which is an application developed by East Anglia University in partnership with La Poste Anglaise. It combines a voice recognition system with a conversation agent (see avatar (c) in Figure 5) to form an intermediary between the employee of "La Poste Anglaise" speaking the English language and the deaf client knowing only British Sign Language (abbreviated as BSL). The signs made by the avatar come from a prior recording of movements actually performed by a person signing in BSL. Recordings are performed by means of a motion capture system placed on the person. This system works for the BSL, but can easily be transposed to other SLs. The motion capture system used for this project, takes into account separately the movements of the hands, the bust, as well as the facial movements. Movements

recorded during capture can be reproduced by a three-dimensional skeleton. This one is then covered with a textured mesh that follows the movements. For this application, 115 short sentences were recorded which covers 90% of La Poste transactions (variable elements such as numbers or days of the week). Concatenating multiple sentences allows the system to generate approximately 370 statements. The advantage of this project is that the animation is fluid and close to human interpretation through the database composed of captured movements by motion capture systems. Also, the TESSA system includes a part recognition of speech to a text, which will then be translated into sign language (Marshall and E. Safar, 2001). The evaluation of the system was performed by the deaf by birth (Cox et al., 2003).

4.1.6 Overview of projects based on Rule-based Approaches

System	Approach	Languages	Direction	Notation	Domain	Evaluation	Human	Bidirectional	Extensible	Animation
Rule-Based MT Systems										
Zardo	I	(A,J,I)SL	to SL	Gloss	Doc-patient					✓
Albuquerque	T	ASL	to SL	Newkirk Not.	Weather					✓
ViSiCAST	T	BSL	to SL	HamNoSys	Narrative		✓			✓
TEAM	I	ASL	to SL	Gloss	n/a					✓
SASL	T	ASL	to SL	Gloss	Doc-patient					✓
ASL W'bench	T	ASL	to SL	Move-Hold	n/a		✓			✓
Spanish SL	T	ASL	to SL	Gloss	Wthr/travel		✓			✓
Multi-path	I&T	ASL	to SL	Part/Cons	n/a	Manual				✓

The table above shows comparison of Related Research based on different criteria, as (Language used, Translation direction, Notation that used in translation, Domain, Evaluation method, Human interaction, Animation support).

4.2 Other Approaches (Example-based, Statistical-based)

In recent years, there have been several research projects related to automatic sign language translation in Europe and in the USA. Systems are based on various types of Algorithms and Machine Translation approach. Like system based on **RBMT** (Rule Based Machine Translation) (Marshall and Eva, 2003) proposed a translating English into British Sign Language using a rule-based approach. Here the grammar was modeled utilizing the HPSG formalism. And the system

is able to translate simple sentences. (Van and Barker 2003) propose another rule-based concept for translating English text to South African Sign Language (SASL). (Huenerfauth, 2004) introduces a rule-based concept for translating English text to American Sign Language (ASL). He concludes that the use of a rule-based approach is more appropriate than the statistical, because large corpora are difficult to obtain. And (San-Segundo et al., 2008) proposed Speech to Sign Language translation system for Spanish in Speech Communication. Another approach based on **SMT** (Statistical Machine Translation), SMT has been the dominant translation paradigm for decades, practical implementations of SMT are generally phrase-based systems (PBMT) which translate sequences of words or phrases where the lengths may differ. (Bauer et al., 1999) proposed a framework for statistical-based Sign Language translation. The authors suggested to translate recognized video based continuous Sign Language to spoken language. (Jan Bungeroth and Hermann Ney 2004) introduce Statistical Sign Language Translation using Distinct corpora for Sign Languages. Additionally, the optimization of the statistical translation process for scarce resources as suggested e.g. by NieBen and Ney (2000) allows for further improvement. And also, (Verónica López-Ludena et al., 2011) introduce Automatic categorization for improving Spanish into Spanish Sign Language machine translation using SFST-based Translation SMT. And give result (SER=19.97%) (PER=17.96%). And (Achraf and Jemni, 2011) proposed Statistical Sign Language Machine Translation: from English written text to American Sign Language Gloss. And (Achraf, 2017) also present “Traduction Automatique En Langue Des Signes Basée Sur Une Approche Statistique” in PhD Thesis, University of Sfax, Tunisian Republic, 2017. Another approach based on **EBML** (Example Based Machine Translation). (Morrissey and A. Way, 2006) presents The Problems of Using Mainstream MT Evaluation Metrics for Sign Language Translation. And Later studies present work that combines EBMT with statistical approaches (Morrissey et al., 2007). The example-based method used in these studies is marker-based chunking. The preferred direction in translation is sign language to spoken language (Morrissey and A. Way et al., 2007) They used a corpus consists of 595 sentences, and best result is between Irish Sign Language (ISL) to English with (BLEU=52.18%). (Morrissey, 2008) present English to ISL resulted about (BLEU=39%), and German to DGS (BLEU=50%). And (Almohimeed and et al., 2011) present Arabic Text to

Arabic Sign Language Translation System for the Deaf and Hearing-Impaired Community with (PER=29.40%) (WER=47.70%). Also (Merve and Ilyas, 2017) introduce Bidirectional Machine Translation Between Turkish And Turkish Sign Language: A DATA-DRIVEN APPROACH using 544 Sentences and the result from TR to TSL: (BLEU=43%) (TER=38%) and from TSL to TR : (BLEU=59%) (TER=47%). In the table below show some result from Written language into sign language translation systems.

Study	Year	TT	Corpus	Dataset	BLEU	SER	PER	WER
Cox et al.	2002		TXT > BSL	< 1000	---	---	---	---
Marshall and Sáfa	2005	HPSG	TXT > BSL	< 1000	---			
Bungeroth and Ney	2004	SMT	TXT > DGS	< 1500	---			
San-Segundo et al.	2008	RBMT	TXT > LSE	< 5000	> 50%	31.60%	27.2%	23.50%
					57.80%			
Verónica López , et al (SFST-based Translation)	2011	SMT	TXT > LSE	4080	> 75%	19.97%	17.96%	
				75% 12.5% 12.5%	78.40%			
Achraf Othman	2017	SMT	TXT <> ASL	119860	< 50%			
				66.67% 22.22% 11.11%	49.73%			
Morrissey et al. (MaTrEx Translation)	2008	SLMT	ISL > TXT	595	> 50%		39.67%	38.48%
				70.25% 9.92% 19.83%	52.18%			
	2008	SLMT	TXT > ISL	595	< 50%		34.33%	46.02%
				70.25% 9.92% 19.83%	38.85%			
A. Almohimeed and et al.	2011	EBMT	TXT > ArSL	203	---		29.40%	46.70%
Merve Simseki and Cicekli	2017	EBMT	TXT > TSL	544	43%	TER = 38%		
	2017	EBMT	TSL > Txt	90.07% -- 9.93%	59%	TER = 47%		

Table above shows the different approaches of sign language translation System and results achieved.

5. Conclusion

In this paper, we have provided a survey of previous and existing techniques for sign language translation. We discussed Rules-based approaches. ViSiCAST system uses CMU Link Parser for analyzing the input text, and then uses Prolog declarative clause grammar rules to convert this linkage output into a Discourse Representation Structure (DRS), and a Head-Driven Phrase Structure Grammar. TEAM system analyzing The input data ,distinguished in syntactic and morphological information. using TAG translation approach with the XTAG grammar for English analysis, and the ASL Workbench uses the Lexico-Functional Grammar formalism. ZARDOZ system use of AI knowledge representation, metaphorical reasoning, and a blackboard system architecture; so, the translation design is very knowledge-based and AI-reasoning heavy. During the analysis stage. Because of the special properties of Sign Language have required the development of new computational linguistic technologies. New sign language symbolic representation systems have been developed for several projects. TEAM uses an extended sign gloss formalism, ViSiCAST converts a similar formalism into HamNoSys-style output, and the ASL Workbench computerizes the Movement-Hold phonological model. The symbolic representation chosen for the sign language output can limit the expressiveness of the eventual output especially for the important None-Manual sign. And discussed other translation methods like EBMT, SMT using several corpora for sign language .

References

- Achraf Othman. (2017). “Traduction Automatique En Langue Des Signes Basée Sur Une Approche Statistique” in PhD Thesis, University of Sfax, Tunisian Republic.
- Achraf Othman, Jemni proposed. (2011). “Statistical Sign Language Machine Translation: from English written text to American Sign Language Gloss”.
- Almohimeed, M. Wald, and R. I. Damper. (2011). “Arabic Text to Arabic Sign Language Translation System for the Deaf and Hearing-Impaired Community,” in Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies.
- Archana S. Ghotkar 1 and Gajanan K. Kharate. (2014). “Study of Vision Based Hand Gesture Recognition Using Indian Sign Language” International Journal on Smart Sensing And Intelligent Systems.
- Bangham, S. J. Cox, Elliott, J. Glauert, I. W. Marshall, S. Rankov and M. Wells. (2000). “Virtual Signing : Capture, Animation, Storage and Transmission - An Overview of the VisiCast Project” chez IEEE seminair on Speech and Language Processing for Disabled and Elderly People.
- Bauer, and et al. (1999). “Video-based continuous sign language recognition using statistical Methods”.
- Cox, M. Lincoln, J. Tryggvason, M. Nakisa, M. Wells, M. Tutt et S. Abbott. (2002). “TESSA, a system to aid communication with deaf people” chez Fifth International ACM SIGCAPH Conference on Assistive Technologies (ASSETS 2002), Edinburgh, Scotland.
- Cox, M. Lincoln, M. Nakisa, M. Wells, M. Tutt et S. Abbott. (2003). “The Development and Evaluation of a Speech to Sign Translation System to Assist Transactions” International Journal of Human Computer Interaction, vol. 16, pp. 141-161.

- DíArmond Speers. (2001). “Representation of American Sign Language for Machine Translation”.
- Elliott, J. R. W. Glauert, V. Jennings et R. Kennaway. (2004).
“SiGML Notation and SiGML Signing Software System” chez Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC).
- Hanke. (2004). “HamNoSys - representing sign language data in language resources and language processing contexts” chez LREC 2004, Workshop proceedings : Representation and processing of sign languages.
- Huenerfauth. (2004). “An Accessibility Motivation for an English-to-ASL Machine Translation System”.
- Jan Bungeroth, Hermann Ney. (2004). “Statistical Sign Language Translation”.
- Jemni et O. El Ghouli. (2008). “A system to make signs using collaborative approach” chez Computers Helping People with Special Needs.
- Klima. (1979). The Signs of Language, MA: Harvard University Press.
- Marshall and E. Safar. (2001). “The Architecture of an English-Text-to-Sign-Languages Translation System” chez Recent Advances in Natural Language Processing (RANLP).
- Marshall, Eva Safar. (2003). “A Prototype Text to British Sign Language Translation System”.
- Mary Elizabeth Bonham. (2015). “English to ASL Gloss Machine Translation”.
- Merve Selcuk-Simsek and Ilyas Cicekli. (2017). “Bidirectional Machine Translation Between Turkish And Turkish Sign Language: A Data-Driven Approach”.

- Morrissey and A. Way. (2006). "Lost in Translation: The Problems of Using Mainstream MT Evaluation Metrics for Sign Language Translation" in Proceedings of the 5th SALTML Workshop on Minority Languages at LREC-06.
- Morrissey and A. Way. (2007). "Joining Hands: Developing A Sign Language Machine Translation System with and for the Deaf Community" in Proceedings of the Conference and Workshop on Assistive Technologies for People with Vision & Hearing Impairments (CVHI).
- Morrissey, A. Way, D. Stein, J. Bungeroth, and H. Ney. (2007). "Combining Data-Driven MT Systems for Improved Sign Language Translation" in Proceedings of Machine Translation Summit XI.
- Morrissey. (2008). "Data-Driven Machine Translation for Sign Languages" PhD Thesis, Dublin City University, Ireland.
- San-Segundo, et al. (2008). "Speech to Sign Language translation system for Spanish" in Speech Communication 50, 1009–1020.
- Stokoe-1960. (2005). "Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf (reprint)" Journal of Deaf Studies and Deaf Education, vol. 10, pp. 3-37.
- Stuart. (2011). M.T. "A Grammar of SignWriting". Thesis in Linguistics, University of North Dakota.
- Van Zijl and Barker. (2003). "South African Sign Language SASL Machine Translation System".
- Veale, Tony, and Alan Conway. (1994). "Cross modal comprehension in ZARDOZ, an English to sign-language translation system." Seventh International Workshop on Natural Language Generation. 249-252.

Verónica López-Ludena et al. (2011). “Automatic categorization for improving Spanish into Spanish Sign Language machine translation”.

Zhao, K. Kipper, W. Schuler, C. Vogler, N. Badler, M. Palmer. (2000). “A machine translation system from English to American Sign Language” chez Proceedings of the 4th Conference of the Association for Machine Translation in the Americas on Envisioning Machine Translation in the Information Future.

An Efficient Ambulance Routing System for Emergency Cases based on Dijkstra's Algorithm, AHP, and GIS

Sayed Ahmed¹, Romani Farid Ibrahim², Hesham A. Hefny³

Abstract. In emergencies cases, the main concern is how to arrive to an incident location in a minimum time. In such cases, we are in the most need to a guide to assets us in determining the nearest ambulance cars to the incident location. In this sense, The current study presents an efficient ambulance routing system based on the integration between Dijkstra algorithm in Artificial Intelligence (AI), spatial data in Geographic Information Systems (GIS), and a decision making approach like the Analytical Hierarchy Process (AHP). This system assists us to find the nearest ambulance cars from an incident location and the best routes to these ambulance cars. It is helpful in terms of reducing transition time, effort, and cost. The proposed system results were compared to those obtained using Google Maps. This comparison showed that is comparable to Google Maps in terms of the travel distance and travel time.

Keywords: AHP, AI, GIS, GPS, sapatial data, Dijkstra algorithm.

I. INTRODUCTION

In a congested area like Grater Cairo Region (GCR), Egypt, finding the desired location is a complicated process, especially in emergency cases. Therefore, it is mandatory to plan and invest wisely in infrastructure for sustainable urban transport due to the vital role that transport plays in urban development by providing people with access to employment, healthcare, education, recreation, and other key services [Ahmed et al., 2017]. Nowadays, the Geographic Information System (GIS) technology is considered as one of the most important research tools in the world in monitoring and emergency management [Abdel Aziz et al., 2016].

GIS is a computerized system that is designed to capture, store, manipulate, analyze, manage, visualize, and present spatial data in a way that a user can easily understand [Shashikiran et al., 2011]. To find out the desired route between any two locations on a road network, either the route that has the minimum length (the shortest route) or the route that has the minimum travel time (the best route) is to be selected.

¹National Authority for Remote Sensing and Space Sciences, Egypt

²Mohammed Bin Naïf Academy for Marine Science and Security Studies, KSA

³Institute of Statistical Studies and Research, Cairo University, Egypt
se.sayedahmed@gmail.com, romanifarid@gmail.com, hehefny@ieee.org

In emergency cases, the best route is preferred, as it takes the minimum travel time to reach an incident location which helps to save people life. There are many factors that affect the selection of the most suitable route, such as road width, velocity limits, traffic volume, etc. These factors have to be carefully evaluated and selected, as they will affect the total time needed for the arrival of the ambulance cars to reach the incident location, and the wrong selection will cause delay in the ambulance cars' arrival.

The shortest path between two nodes "S" and "D" on a network can be defined as the directed simple path from node "S" to node "D", provided that no other path has a lower weight [Kong et al., 2011]. The majority of applications solve the shortest path problem using only the path's length as a weight, but in emergencies it will not provide the optimal results. there are various factors that must be taken into account in order to minimize delays as much as possible. Analyzing, comparing, prioritizing, and weighting these factors represent a big challenge.

AHP, developed by Saaty (1980), is a Multi Criteria Decision Making (MCDM) approach that is used in complex decision-making problems. It follows a multi-level hierarchical structure with a final goal and the possible alternatives to achieve this goal [Mali et al., 2013]. AHP not only helps the decision makers to take the appropriate decision, but also to find the one that best fit for their needs and understanding to the problem [Yang and Li, 2010]. It is widely used in a variety of decision making situations in various fields such as government, healthcare, industry, education, and business [Yang and Li, 2010]. In this research, we have used the AHP method to analyze and compare the eight impedance factors that affect the selection of the best route, and give a weight (priority) for each factor.

The rest of this paper is organized as follows; Section 2 discusses some previous related works , Section 3 discusses the proposed methodology of the research, Section 4 discusses the experimental setups and results obtained from this study, and finally Section 5 concludes the study and provides an outlook towards the future works.

II. RELATED WORKS

In [Ahmed et al., 2017], the authors applied an enhanced GIS-based network analysis on the GCR road network. They focused on finding the best route between two locations on the GCR road network and finding the closest healthcare service providers to an accident location based on the travel time. In their proposed methodology, the authors have integrated historical traffic data to be used in the analysis. They

have used the Dijkstra's routing algorithm built into the ArcGIS Network Analyst Extension for performing the network analysis. The proposed method has improved the travel time by 22%, depending on the travel distances.

In [Bozyigit et al., 2017], the authors improved Dijkstra's algorithm to overcome its two shortcomings (walking distance and ignoring number of transfers) by developing a penalty function. The proposed penalty function adds a small penalty cost to the cost of the path in case of stated shortcomings. Thus, the number of transfers and the number of repeated walking were minimized by slightly increasing the distance of the proposed path in the new route selection method. The improved algorithm outperformed the traditional algorithm in terms of the average number of transfers and the average walking distance, but it had the worst running time.

In [Risald et al., 2017], the authors integrated Dijkstra's algorithm with Floyd-Warshall algorithm in a best route selection method. Floyd-Warshall algorithm was used to compute the distance to the nearest hospital, and then the minimum travel time route to this hospital was determined using Dijkstra's algorithm. Factors taken into account were the distance to the nearest hospital and the traffic flow density. If some unexpected problems occurred on the way to the nearest hospital due to unexpected factors such as traffic incidents, the system moves to other nearby hospital. Integrating Floyd-Warshall algorithm with Dijkstra's algorithm masked the defeats of each algorithm and speeded up the search time by reducing unnecessary decision makings.

In [Kukadapwar and Parbat, 2016], the authors used AHP technique and the classical Dijkstra's algorithm to find optimal path in real urban road network. The classical Dijkstra's algorithm was used to select the shortest paths between any two stop points, then the AHP method was used to select the optimal path among them. The authors used the AHP method to analyze, compare, and weight some factors that affect the selection of the optimal path. The factors considered were: distance, time, traffic volume, road width, the number of intersections, parking, and road's encroachment.

In [Gubara et al., 2014], the authors developed an emergency response system based on GIS to obtain the best route from an incident location to the nearest healthcare service provider. The best route was modeled based on the length of the route only. The authors did not consider any factors other than the path length.

In [Kai et al., 2014], the authors developed an emergency response system based on GIS and Dijkstra's algorithm. The proposed system has integrated web services and Asynchronous JavaScript and XML

(Ajax) technologies to build a web-based interface for finding the best routes from the specialized response team stations locations to an incident location. Their proposed system provided the optimal route depending only on the length of the route without considering any other factors that impose the delay on the travelling time.

In [Dabhade et al., 2015], the authors have used Dijkstra's algorithm and the ArcGIS software to find the shortest path to reach a specialized hospital in Aurangabad city, Maharashtra State, India.

III. PROPOSED METHODOLOGY

The proposed system was built to assist in making a decision for the nearest ambulance cars to an incident location and finding the best routes to these facilities by using GIS and an improved Dijkstra algorithm. The architecture of the proposed system has many stages as shown in Fig.1.

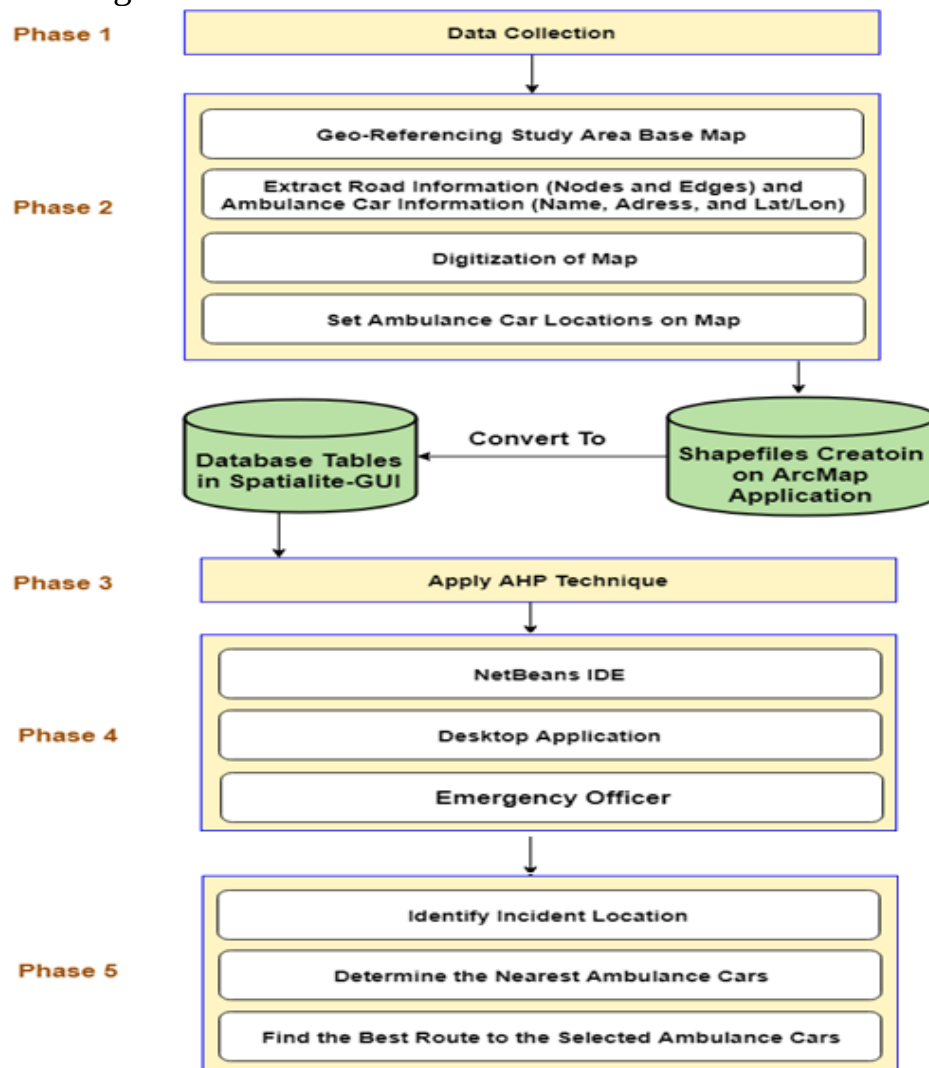


Fig. 1. The proposed system stages.

Many software were used to implement the proposed system, these software are:-

- ArcGIS program.
- Spatialite – GUI (a GUI tool for Spatialite).
- NetBeans IDE version 8.1.

In addition to:-

- Java Programming Language.

Phase 1: Data Collection

The first stage of the proposed system is to collect the required data. The required data for the proposed system can be categorized in two main categories: spatial data and attribute data.

- **Spatial Data:** The data required for the system are the locations for road nodes (coordinates latitude and longitude), edges of road network, and the ambulance cars locations. This spatial data were obtained through the process of digitizing base map of specific area and the locations coordinates for the ambulance cars were collected using GPS.
- **Attribute Data:** Ambulance cars data collected using GPS. This data includes ID, name, and address. The attribute data of roads network include two parts; the first part is the node features such as node name and node ID, and the second part is the edge features such as edge ID, name, the start node of the edge, the end node of the edge, length and other edges information.

Phase 2: GIS Utilization

This phase is utilizing GIS capabilities for preparing the geographic data necessary for the analysis performed in the proposed system. This phase has many steps which are discussed in the following subsections.

Geo-Referencing Study Area Base Map

The study area of the proposed system is the roads network of GCR. In this step, a geo-referencing process for the data obtained in the first phase is being performed. Geo-Referencing process allows the registration of the digitized top sheet on the earth's surface [Dabhade et al., 2015]. It is considered a very critical stage as it affects the accuracy of the obtained data. The study area base map and roads network are shown in Fig. 2 and Fig. 3. A base map can be obtained as an image from many resources such as Google earth, Google maps, Bing maps, OSM, or from any other sources that provide map.

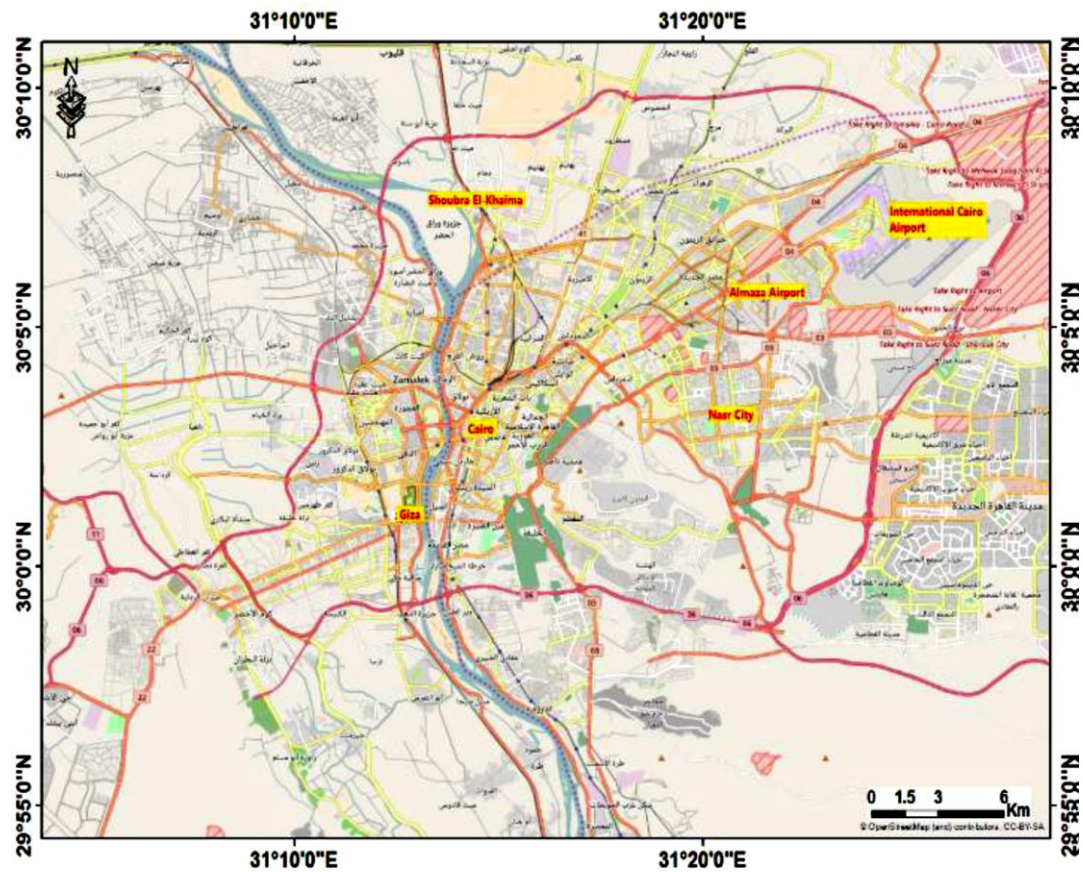


Fig. 2. GCR base map (from OSM).

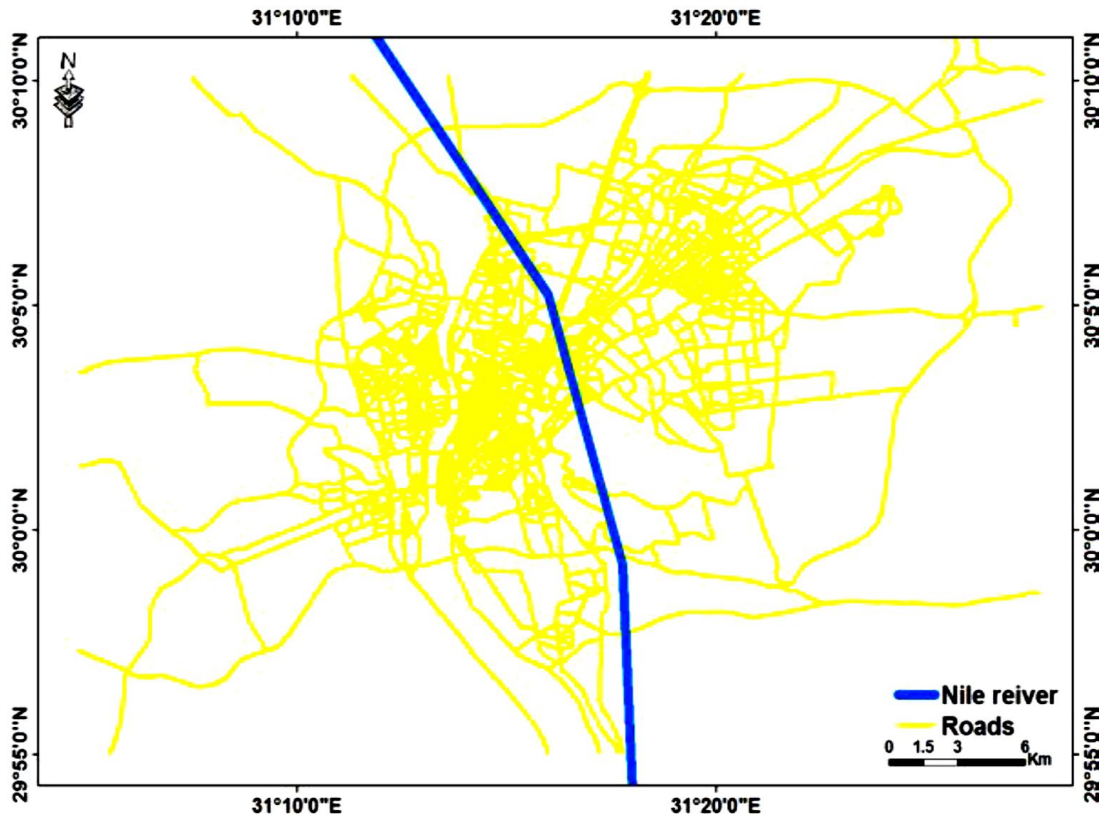


Fig. 3. GCR roads network.

Extract Roads Information (Nodes and Edges) and Ambulance Car Information (Name, Address, and Lat/Lon)

In this step we extract the information of the roads network and the ambulance car that were obtained from OSM. OSM data are available for commercial and public organizations. It has open sources map, which means that the user can use these information and update or correct the data on map [Arruda, 2011]. The study area roads network consists of 3970 nodes and 11140 edges. In addition, the ambulance car data obtained has a total of 14 locations that cover the study area. The steps required for obtaining the desired format of data for being used in the implemented system are shown in Fig. 3 [Muhsen and Hassan, 2018].

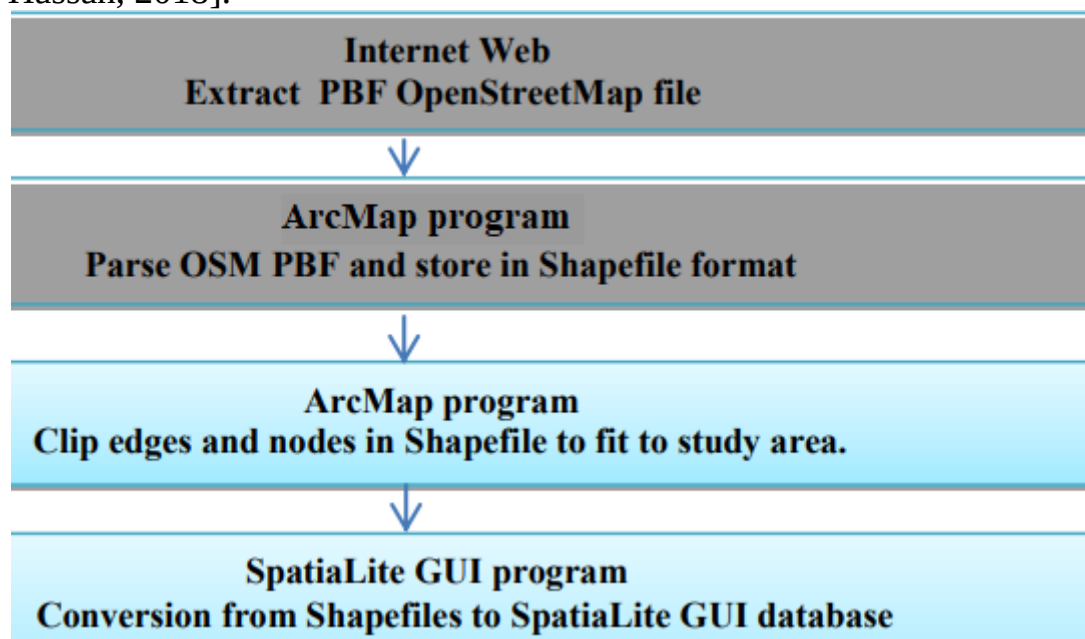


Fig. 4. The steps of processing data set.

Digitization of Map

It is the operation of changing paper map properties to digital formula [Dabhade et al., 2015]. This process produces the geographic data in vector data format such as line, point, polygon, etc. as coordinates (latitude and longitude) to make it easy to store and visualize this information on computers [Muhsen and Hassan, 2018].

Set Ambulance Cars Locations on Map

The coordinates (lat/lon) of the ambulance cars locations were taken using Global Positioning System (GPS), then these coordinates are specified on map as points feature. Fourteen ambulance cars locations in GCR were studied through getting coordinates for their locations by GPS.

These points are stored in shapefile named ambulances and converted into a Spatialite-GUI database table with the same name.

Phase3: Apply AHP Technique

In this phase, AHP method was used to analyze and compare eight impedance factors and calculate the overall priority of each factor through building comparison matrix. Then the classical Dijkstra's algorithm was enhanced by modifying its weight factor to represent the total weight for each road in the road network, and applied on a weighted road network to give the best route between any two given points [Ahmed et al., 2018]. The steps of this phase are discussed in the following subsections.

Build AHP hierarchy

The first step in AHP technique is to build the AHP hierarchy, i.e., representing the problem as a hierarchy in which the overall goal is the root and the set of criteria are branching from that root. In our methodology, we have selected eight impedance factors that affect the selection of the best routes. The eight impedance factors considered to achieve the goal are represented as a set of criteria. The impedance factors and the AHP hierarchy are shown in Fig. 5 [Ahmed et al., 2018].

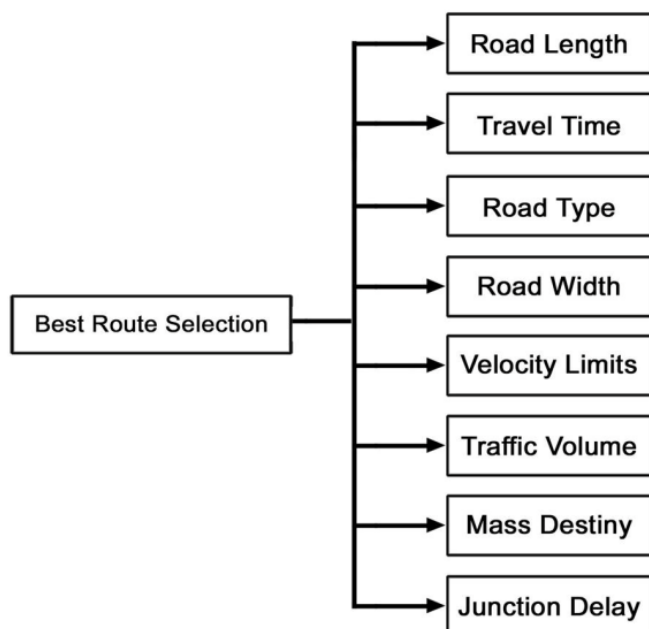


Fig. 5. The AHP hierarchy model of the proposed methodology.

Make Pairwise Comparisons

In AHP process, the pairwise comparison has two roles. First, it is made to judge the importance of one criterion over another. Second, it

is made to obtain the overall weight vector for all criteria. Judging the importance of one criterion over another is a subjective issue that depends on one's perception. Table 1 lists the quantitative scales for the pairwise comparisons [Ahmed et al., 2018].

Table 1. Scales in pair wise comparisons

Importance Scale	Definition
1	Equally Importance
3	Moderate Importance
5	Strong Importance
7	Extreme Importance
9	Extremely More Importance
2, 4, 6, 8	Intermediate values between Adjacent Scale Values

The number of pairwise comparisons that will be made is determined according to Eq. (1)

$$\text{No. of Comparisons} = n \times (n-1)/2 \quad (1)$$

Such that, n is the number of criteria. Therefore, according to Eq.(1), the number of pair wise comparisons is 28 comparison.

Build Pairwise Comparison Matrix

The output of the pairwise comparison made in the previous step is a comparison matrix which gives the importance ratios for each pair of alternatives. The Comparison Matrix is an $n \times n$ matrix, such that n is the number of criteria. The values in the comparison matrix are all positive. The main diagonal of this matrix is always 1, which represent the value of comparing each criterion to itself. The values above the main diagonal can be expressed as a_{ij} , which indicates the importance of i^{th} criterion over j^{th} criterion. The values below the main diagonal are the inverse of the values above the main diagonal and can be expressed as a_{ji} and can be calculated as $1/a_{ij}$.

Build the Normalized Comparison Matrix

The overall weight of each factor is calculated by normalizing the comparison matrix of the previous step using the geometric mean method given in Eq. (2).

$$r_i = \prod_{j=1}^n (a_{ij})^{1/n} \quad (2)$$

Where a_{ij} ($i, j=1 \dots n$) are the comparison values in the pairwise comparison matrix and n is number of alternatives.

Compute the Overall Weight Vector

After computing the normalized comparison matrix, the overall weight vector can be calculated as shown in Eq. (3)

$$w_i = r_i / \sum_j r_j \quad (3)$$

Compute the Consistency Ratio C.R.

The last step in AHP is to compute the consistency ratio of the pairwise comparison matrix. Consistency Ratio (C.R.) is a measure of the consistency of the comparison matrix according to your judge and perception. The Consistency Ratio can be calculated as shown in Eq. (4).

$$C.R. = C.I. / R.I. \quad (4)$$

Such that, C.I. is the Consistency Index and R.I. is a Random Index. Saaty gave C.I. as a deviation or degree of consistency, and can be calculated as shown in Eq. (5) [Ahmed et al., 2018].

$$C.I. = \frac{(\lambda_{max} - n)}{n-1} \quad (5)$$

Such that, λ_{max} is the maximum Eigen value of comparison matrix and n is the number of alternatives. Saaty proposed that we can use this index by comparing it with the appropriate one. The appropriate consistency index is called Random Consistency Index (R.I.). The Random Consistency Index (R.I) is shown in Table 2.

Table 2. Values of the Random Consistency Index (RI) for small problems [Ahmed et al., 2018]

n	1	2	3	4	5	6	7	8
R.I	0	0	0.58	0.90	1.12	1.24	1.32	1.41

Check the Comparison Matrix Consistency

If $C.R. < 10\%$, then the pairwise comparison matrix is consistent, and in this case the final decision matrix and the overall weight vector were

produced, otherwise, we have to repeat the pairwise comparison, i.e., repeat the methodology steps from 3.2 to 3.7. The final decision matrix and the overall weight vector are shown in Table 3 and Table 4 respectively.

Table 3. The final Decision Matrix

	RL	TT	RT	RW	VL	TV	MD	JD
RL	1	0.11	0.5	0.14	0.11	0.11	0.11	2.0
TT	9.0	1	9.0	3.0	2.0	1.0	1.0	9.0
RT	2.0	0.11	1	1.0	0.14	0.11	0.11	3.0
RW	7.0	0.33	1.0	1	0.5	1.0	1.0	7.0
VL	9.0	0.5	7.0	2.0	1	1.0	1.0	7.0
TV	9.0	1.0	9.0	1.0	1.0	1	1.0	7.0
MD	9.0	1.0	9.0	1.0	1.0	1.0	1	7.0
JD	0.5	0.11	0.33	0.14	0.14	0.14	0.14	1

Table 4. The overall weight vector

Criteria	Weight	Rank
RL	2.2%	7
TT	24.2%	1
RT	4.1%	6
RW	11.6%	5
VL	17.9%	4
TV	19.0%	2
MD	19.0%	2
JD	2.0%	8

Combine AHP with Dijkstra's Algorithm

After generating the final weight vector, we will use it to compute an overall weight for each route in the routes network, and use this general weight to modify the weight factor of the classical Dijkstra's algorithm. Due to the diversity of measurement units for the impedance factors, it is impossible to express these in one factor, so a normalization process for each impedance factor was applied. The objective of the normalization process is to eliminate the units of measurement for the impedance factors and generate values between 0 and 1, which can be easily combined in a single value. Not all impedance factors will be normalized in the same way. Some of them need to be maximized to select the best route, such as Velocity Limits, Road Type and Road Width. The other impedance factors need to be minimized to select the best route, such as Road Length, Initial Travel Time, Traffic Volume, Mass Destiny and Junction Delay. For those

impedance factors that need to be maximized, the normalization formula will be as shown in Eq. (6).

$$NF_n = 1 - (F_n - \text{Min}(F_n) / \text{Max}(F_n) - \text{Min}(F_n)) \quad (6)$$

Where:

NF_n = Normalized Factor F_n

F_n = Old value of Factor F_n

$\text{Min}(F_n)$ = Minimum value of the Factor F_n

$\text{Max}(F_n)$ = Maximum value of the Factor F_n

$n = 1 \dots 8$.

For those impedance factors that need to be minimized, the normalization formula will be as shown in Eq. (7).

$$NF_n = (F_n - \text{Min}(F_n) / \text{Max}(F_n) - \text{Min}(F_n)) \quad (7)$$

Finally, the total weight for each route in the routes network will be calculated as shown in Eq. (8).

$$W_T = \sum_{n=1}^8 (NF_n \times W_n) \quad (8)$$

Where:

W_T = the total weight for each route in the routes network in terms of the eight impedance factors.

NF_n = Normalized Factor F_n .

W_n = the corresponding weight of the factor F_n in the overall weight vector.

After the general weight for each route has been calculated, it will be used as the weight value for the classical Dijkstra's Algorithm.

Phase4: Implement the Proposed System

This phase is related to the implementation of the desktop application that enables the user to interact with components of the application, display the study area roads network, set the incident location, search for the nearest ambulance car locations, display the search result, and to find the best route to each selected ambulance car.

NetBeans IDE

The developers of Java-based desktop applications need an efficient development environment to build their applications. NetBeans IDE is one applicable choice. It provide developers with the tools and features

that facilitate their work.

The Desktop Application

In the Main window of the desktop application the map and GUI components are initialized.

Table 5 demonstrate the main features and functions of the implemented application.

Table 5. The main features of the implemented application

Feature	Description
Draw Graph	Draw the GCR roads network (Nodes and Edges) on the main window.
Load Graph Nodes	Populate the source nodes drop-down list with the oads network nodes.
Add Source	Add the incident location (source node) on the roads network.
Nearest Ambulances	Seach for the nearest five ambulanc car locations to the incident location and draw them on the roads network
Find Route	Calculate the best route from the incident location to each of the nearset five ambulance car locations and draw them on the roads network
Reset	Resetting the application to its initial state (Only the roads network is displayed)

The Emergency Officer

The user who will interact with the application is the emergency officer. He is responsible for receiving the emergency call, determine the incident location, searching for the nearest ambulance car locations to the incident location, finding the best route from each location to the incident location, and inform the selected ambulance car with all this information to reach the incident location. The application main window is shown in Fig. 6. Fig. 7 demonstrate all of features listed in Table 6.

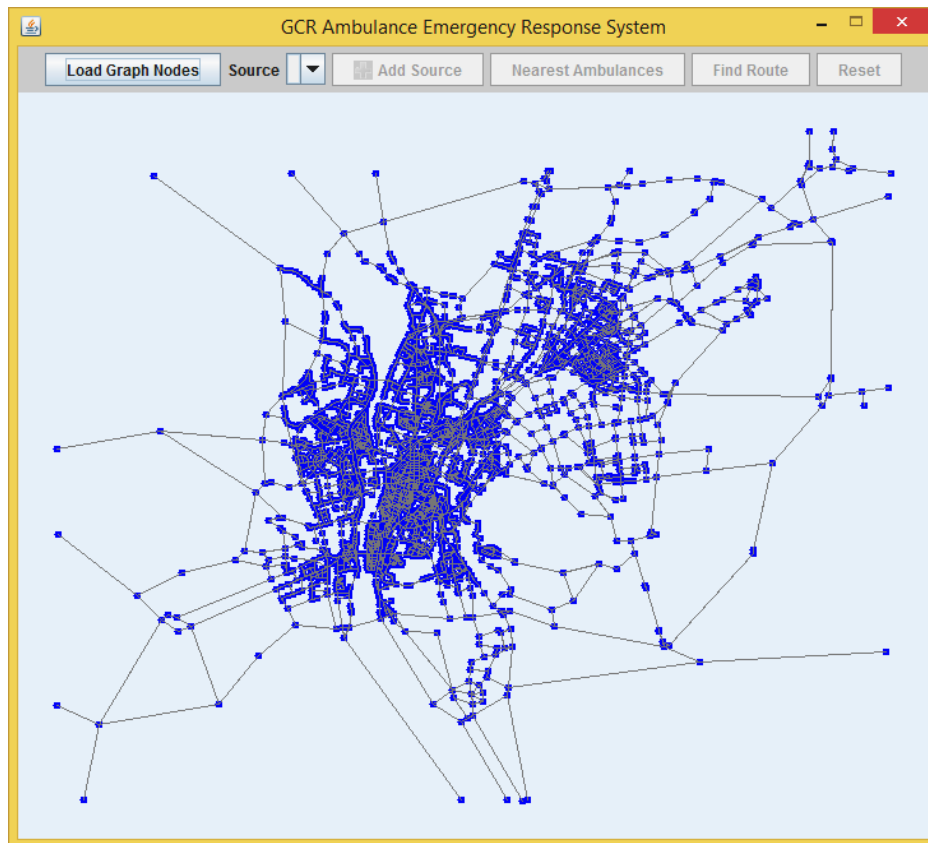


Fig. 6.The application main window.

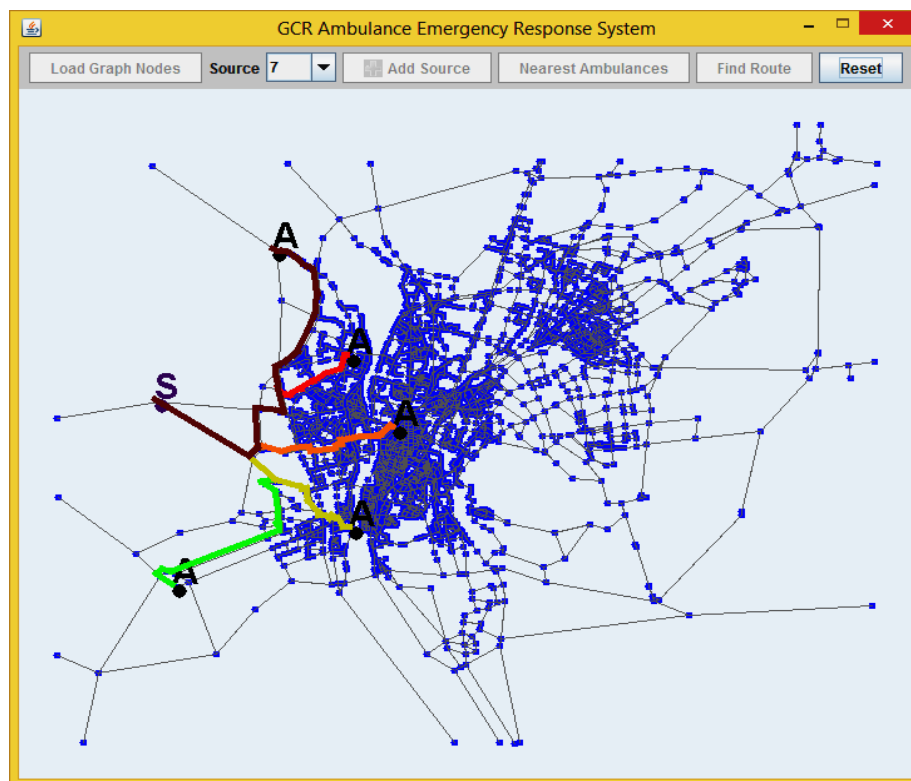


Fig. 7.The main features of the system.

Phase5: Test the Proposed System

This phase is very important because it's the core of the proposed system. The proposed system depends on the the integration between GIS and Dijkstra algorithm. The steps of this phase are discussed in the following subsections.

Identify Incident Location

The first step in this phase is to determine the coordinates location (lat/lon) of the incident. This location represents the source node needed for Dijkstra's algorithm to start. For the simplicity of calculation, we assumed that the incident location will be identical to the location of some roads' node.

Determine the Nearest Ambulance Cars

The second step in this phase is to determine the nearest ambulance cars locations to the incident location. This is performed by calculating the direct distance from the incident location to each ambulance car's location using the haversine method. The haversine formula determines the great-circle distance between two points on a sphere given their latitudes and longitudes, i.e. the shortest direct distance between two points over the earth's surface [Alam et al., 2016]. The haversin formula is shown in Eq. (9), Eq. (10), and Eq. (11) respectively.

$$a = \sin^2(\Delta\phi/2) + \cos \phi_1 \cdot \cos \phi_2 \cdot \sin^2(\Delta\lambda/2) \quad (9)$$

$$c = 2 \cdot \operatorname{atan}^2(\sqrt{a}, \sqrt{1-a}) \quad (10)$$

$$d = R \cdot c \quad (11)$$

Where:

ϕ_1, ϕ_2 : latitude of first point and second point (in radian measure)

λ_1, λ_2 : longitude of first point and second point (in radian measure)

R: the earth's radius (mean radius = 6,371km)

$\Delta\phi$: $\phi_1 - \phi_2$

$\Delta\lambda$: $\lambda_1 - \lambda_2$

d: distance

The haversine formula is applied between the incident location and each of the sixteen ambulance car locations, then the five ambulance car locations that has the least distance are being selected and displayed on the map.

Find the Best Routes to the Selected Ambulance Cars

The results of the previous steps in this phase are: the incident location and the five nearest ambulance car locations. The incident location is used as the source point for the enhanced Dijkstra's algorithm, and the five nearest ambulance car locations are used as the algorithm's multiple destination points. Then, the algorithm calculates the best route to every selected ambulance car.

IV. RESULTS AND DISCUSSION

For testing the proposed system in a near real-time GIS usage, experiments and comparisons were implemented to ensure that the proposed system works properly.

AHP Results Consistency

To judge the consistency of results obtained from the AHP method, we need to calculate the C.R. value. If $C.R. < 0.1$, then the AHP results are consistent. In this research, $C.R. = 0.058$, i.e., < 0.1 , which means high consistency for the AHP results. In addition, was produced after 5 iterations of rebuilding the pairwise comparison matrix. Also, the final Principle Eigenvalue computed in this methodology is $\lambda_{max} = 8.566$ and the Eigenvector solution resulted after 5 iterations.

The Improved Dijkstra's Algorithm

Using the implemented desktop application, we have tested both the modified Dijkstra's algorithms on the routes network of GCR. In emergencies cases, the emergency officer needs to get the necessary information about the case. The most important information is the incident location. Based on the incident location, the system will search for the nearest ambulance cars locations. Then, the system displays the nearest five ambulance cars location. Finally, the suggested route for each location is plotted on the map along with a navigation report that shows the navigation details from the incident location to each selected ambulance car location. These information are the incident location (lat/lon), the ambulance car location (lat/lon), the address of the ambulance car location, the travel distance, and the travel time.

Fig. 8 shows an example scenario of emergency case. The incident location is selected from the "Source" drop-down list and added on the road network map using the button "Add Source". We have assumed that the incident location is equal to the location of the road network vertex 1696. The incident location is represented on the road network map with the capital letter "S". Then, using the button "Nearest

Ambulances” the nearest (direct distance) five ambulance cars are selected and located on the map expressed as “A0”, “A1”, ..., “A4” respectively. Then, using the “Find Route” button, the best route for each selected ambulance is calculated and plotted on the map. Finally, a pop-up window including the navigation details for each route is displayed as shown in Fig. 9.

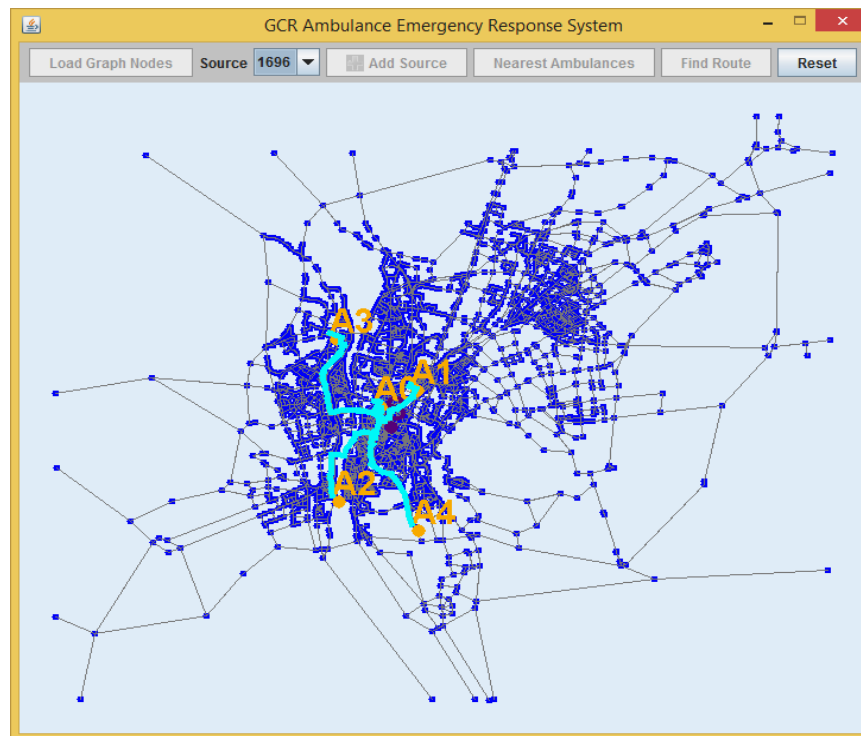


Fig. 8.Example for emergency case scenario.

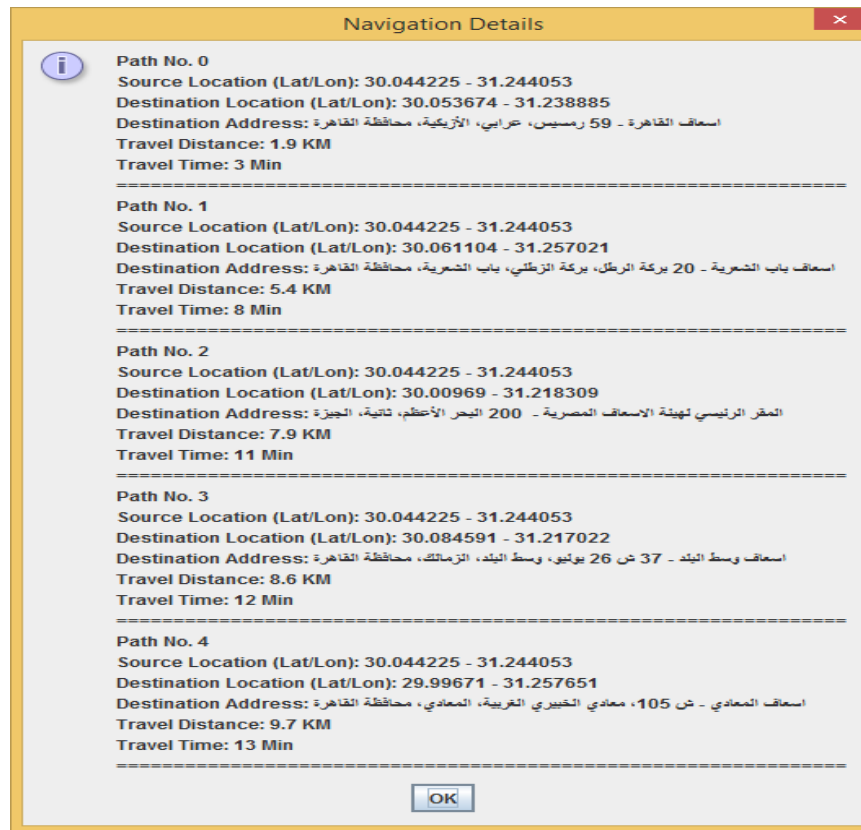


Fig. 9.Navigation details of the previous scenraio.

Evaluation

To ensure the efficiency of our proposed system, we have randomly selected 12 start node to represent the incident locations and calculated the best routes to the nearest 3 ambulance car locations to each incident location using our enhanced Dijkstra's algorithm. We have compared the results of our enhanced Dijkstra's algorithm with the results of getting the best route between each incident location its 3 nearest ambulance cars locations using Google Maps. Table 6 and Table 7 show the results of this comparison.

Table 6.The Results for distance between the source and destination using the proposed system and Google Maps.

No	Path		Google Map (KM)	Propsed System (KM)
	Source	Destination		
1	شارع البلقاسي	اسعاف القاهرة	1.8	1.9
2	شارع البلقاسي	اسعاف باب الشعريه	3.4	5.4
3	شارع البلقاسي	هيئة الاسعاف المصرية	6.8	7.9
4	شارع السيد الببلاوي	اسعاف القاهرة	3.7	4
5	شارع السيد الببلاوي	اسعاف باب الشعريه	4.8	5.6

6	شارع السيد الببلاوي	اسعاف المعادي	5.5	7
7	شارع المطار	اسعاف وسط البلد	4.9	3.7
8	شارع المطار	اسعاف القاهرة	7.8	9.9
9	شارع المطار	اسعاف العبادية	20	11.3
10	شارع رشيد	اسعاف وسط البلد	3.4	4
11	شارع رشيد	اسعاف القاهرة	6.4	5.1
12	شارع رشيد	اسعاف باب الشعرية	9	8.9
13	شارع عطبرة	اسعاف وسط البلد	3	3.6
14	شارع عطبرة	اسعاف القاهرة	6.1	5.5
15	شارع عطبرة	اسعاف باب الشعرية	8.7	9.3
16	طريق امبابة القناطر الزراعي	اسعاف العبادية	2.1	10.9
17	طريق امبابة القناطر الزراعي	اسعاف وسط البلد	5.7	5.6
18	طريق امبابة القناطر الزراعي	اسعاف ابو الغيط	14.2	8.3
19	الطريق الدائري الوراق	اسعاف العبادية	2.3	11.2
20	الطريق الدائري الوراق	اسعاف وسط البلد	6.5	6
21	الطريق الدائري الوراق	اسعاف ابو الغيط	14.6	8
22	شارع كورنيش النيل	اسعاف العبادية	1.8	1.9
23	شارع كورنيش النيل	اسعاف وسط البلد	6.8	5.9
24	شارع كورنيش النيل	اسعاف ابو الغيط	14.6	8.6
25	شارع العروبة	اسعاف وسط البلد	2.1	2.4
26	شارع العروبة	اسعاف العبادية	15.3	9.9
27	شارع العروبة	اسعاف ابو الغيط	10.9	11.9
28	شارع القاضي	اسعاف وسط البلد	2.2	1.8
29	شارع القاضي	اسعاف القاهرة	6.1	6.1
30	شارع القاضي	اسعاف باب الشعرية	8.7	9.9
31	شارع المواردي	اسعاف القاهرة	3.1	4.4
32	شارع المواردي	هيئة الاسعاف المصرية	6.7	5.2
33	شارع المواردي	اسعاف المعادي	6	7.6
34	شارع منية السرج	اسعاف وسط البلد	5	4.3
35	شارع منية السرج	اسعاف باب الشعرية	7.3	8.3
36	شارع منية السرج	اسعاف القاهرة	7.5	6.2

Table 7.Results for time between the source and destination using the proposed system and Google Maps.

No	Path		Google Map (MIN)	Proposed System (MIN)
	Source	Destination		
1	شارع البلقاسي	اسعاف القاهرة	10	3
2	شارع البلقاسي	اسعاف باب الشعرية	18	8
3	شارع البلقاسي	هيئة الاسعاف المصرية	19	11

4	شارع السيد الببلاوي	اسعاف القاهرة	18	6
5	شارع السيد الببلاوي	اسعاف باب الشعرية	29	8
6	شارع السيد الببلاوي	اسعاف المعادي	18	9
7	شارع المطار	اسعاف وسط البلد	25	6
8	شارع المطار	اسعاف القاهرة	22	13
9	شارع المطار	اسعاف العبادية	32	14
10	شارع رشيد	اسعاف وسط البلد	21	6
11	شارع رشيد	اسعاف القاهرة	18	7
12	شارع رشيد	اسعاف باب الشعرية	27	12
13	شارع عطبرة	اسعاف وسط البلد	19	6
14	شارع عطبرة	اسعاف القاهرة	20	8
15	شارع عطبرة	اسعاف باب الشعرية	28	13
16	طريق امبابة القناطر الزراعي	اسعاف العبادية	6	12
17	طريق امبابة القناطر الزراعي	اسعاف وسط البلد	28	7
18	طريق امبابة القناطر الزراعي	اسعاف ابو الغيط	28	10
19	الطريق الدائري الوراق	اسعاف العبادية	6	13
20	الطريق الدائري الوراق	اسعاف وسط البلد	25	8
21	الطريق الدائري الوراق	اسعاف ابو الغيط	26	9
22	شارع كورنيش النيل	اسعاف العبادية	5	3
23	شارع كورنيش النيل	اسعاف وسط البلد	27	8
24	شارع كورنيش النيل	اسعاف ابو الغيط	27	10
25	شارع العروبة	اسعاف وسط البلد	17	4
26	شارع العروبة	اسعاف العبادية	24	12
27	شارع العروبة	اسعاف ابو الغيط	22	14
28	شارع القاضي	اسعاف وسط البلد	12	3
29	شارع القاضي	اسعاف القاهرة	20	9
30	شارع القاضي	اسعاف باب الشعرية	26	13
31	شارع المواردي	اسعاف القاهرة	6	7
32	شارع المواردي	هيئة الاسعاف المصرية	13	8
33	شارع المواردي	اسعاف المعادي	17	10
34	شارع منية السرج	اسعاف وسط البلد	18	6
35	شارع منية السرج	اسعاف باب الشعرية	22	11
36	شارع منية السرج	اسعاف القاهرة	21	9

From Table 6 and Table 7 it can be noted that there are differences between the results obtained by the improved Dijkstra algorithm in the proposed system and those of the Google maps. This is due to the fact that the enhanced algorithm takes the sub-roads in its path in addition to main roads when calculating the distance between the current location and the nearest ambulance cars locations, while the Google map often uses the main roads to find the distance. Also, Google Maps

travel time depends on real-time traffic information, so this time is more accurate than the travel time obtained by the improved Dijkstra's algorithm in our proposed system which incorporate historical traffic information.

From Table 6, it can be noted that there are three types of results:

- Proposed system results are smaller than Google map (17 path highlighted in Green)
- Proposed system results are greater than Google map (18 path highlighted in Red)
- Proposed system results are the same as Google map (1 path highlighted in White)

Also, From Table 7, it can be noted that there are two types of results:

- Proposed system results are smaller than Google map (33 path highlighted in Green)
- Proposed system results are greater than Google map (3 path highlighted in Red)

V. CONCLUSION

Recently the spatial data are very important in our daily life to contribute in making decisions related to GIS. The essential objective of the current study was to build an efficient ambulance routing system based on the integration between an enhanced Dijkstra algorithm in artificial intelligence, spatial data in GIS, and a decision making approach like AHP.

This system is considered an excellent guide for emergencies officers or patient to benefit from its services in a very short time, cost, and effort by giving decisions for the nearest ambulance car to an incident location, the best routes to reach these ambulance cars, and all information related to this facility. We have compared the proposed system results with those obtained using Google Maps. This comparison showed that is comparable to Google Maps in terms of the travel distance and travel time.

VI. REFERENCES

Ahmed, S., Ibrahim, R. F., and Hefny, H. A., "GIS-Based Network Analysis for the Roads Network of the Greater Cairo Area", In: Proc. of 2nd International Conference on Applied Research in Computer Science and Engineering, 2017.

Abdel Aziz, A. E., Abdalla, R. F., Gism, N. A., Mohammed, A. E., and Khidir, S. E., "Route Network Analysis in Khartoum City", Journal of Engineering and Computer Science, Vol. 17, No. 1, pp.50-57, 2016.

Shashikiran, V., Kumar, T. S., Kumar, N. S., Venkateswaran, V., and Balaji, S., "Dynamic road traffic management based on krushkal's algorithm", In: Proc. of International Conference on Recent Trends in Information Technology, pp.200-204, 2011.

Kong, D., Liang, Y., Ma, X., and Zhang, L., "Improvement and Realization of Dijkstra Algorithm in GIS of Depot", In: Proc. of International Conference in Control, Automation and Systems Engineering, pp.1-4, 2011.

Mali, V., Rao, M., and Mantha, S. S., "AHP Driven GIS Based Emergency Routing in Disaster Management", In: Proc. of Communications in Computer and Information Science, pp.237-248, 2013.

Yang, S. and Li, C., "An enhanced routing method with Dijkstra algorithm and AHP analysis in GIS-based emergency plan", In Proc. of 18th International Conference on Geoinformatics, pp.1-6, 2010.

Bozyigit, A., Alankus, G., and Nasiboglu, E., "Public Transport Route Planning: Modified Dijkstra's Algorithm", In: Proc. of the 2nd International Conference on Computer Science and Engineering, pp.502–505, 2017.

Risald, Mirino, A. E., and Suyoto, "Best Routes Selection Using Dijkstra and Floyd-Warshall Algorithm", In: Proc. of International Conference on Information & Communication Technology and System, pp.155–158, 2017.

Kukadapwar, S. and Parbat, D., "Estimation of Optimal Path on Urban Road Networks Using AHP Algorithm", International Journal for Traffic and Transport Engineering, Vol.6, No.1, pp.13-24, 2016.

Gubara, A., Amasha, A., Ahmed, Z., and El Ghazali, S., "Decision Support System Network Analysis for Emergency Applications", In: Proc. of 9th International Conference on Informatics and Systems, pp.40-44, 2014.

Kai, N., Ting, Z. Y., and peng, M. Y., "Shortest Path Analysis Based on Dijkstra's Algorithm in Emergency Response System", TELKOMNIKA Indonesian Journal of Electrical Engineering, Vol.12, No.5, pp.3476-3482, 2014.

Dabhade, A., Kale, K. V., and Gedam, Y., "Network Analysis for Finding Shortest Path in Hospital Information System", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.5, No.7, pp.618-623, 2015.

Arruda, C. D., "Developing a Pedestrian Route Network Service (PRNS)", M.Sc. thesis, Department of Earth and Ecosystem Sciences, Physical Geography and Ecosystems Analysis, Lund University, 2011.

Muhsen, D. K. and Hassan, R. F., "The Intelligent SDSS Mobile Application to Find the Closest Hospital Using the Improved Dijkstra Algorithm with GIS and GPS", Al-Ma'mon College Journal, No.31, pp.239-266, 2018.

Ahmed, S., Ibrahim, R. F., and Hefny, H. A., "Mobile-based Routes Network Analysis for Emergency Response Using an Enhanced Dijkstra's Algorithm and AHP", International Journal of Intelligent Engineering and Systems, Vol.11, No.6, pp.252-260, 2018.

Alam, C. N., Manaf, K., Atmadja, A. R., and Aurum, D. K., "Implementation of Haversine Formula for Counting Event Visitor in The Radius Based on Android Application ", In: Proc. of 4th International Conference on Cyber and IT Service Management, pp.1-6, 2016.

Survey on Social Network Analysis

M. M. Abo Khedra¹

A. A. Abd EL-Aziz²

Hesham A. Hefny³

Abstract- Social Network Analysis determines the relationships between people, organizations, groups or entities and the flow of information between them. A literature survey is done for social network analysis and the prediction of links that could be done in future between the different entities within the network. Many use cases are shown here for applying Social Network Analysis in different industries. This survey introduces different approaches to anticipating the connection between nodes in the social network. It will also support future research and development work as well as raising the awareness for the presented approaches.

Keywords— *SNA, Influencer, Acquisition, Community Detection, Link Prediction*

1 INTRODUCTION

A Social network is a Public structure consists of individuals called “nodes” that are tied by one or more types of links, such as friendship, kinship, financial exchange, common interest, dislike, or relationships of beliefs or knowledge. Social Network Analysis (SNA) is a process of investigating social structures using the network, that characterizes its structure in terms of nodes (people, individual actors or things inside the network) and edges (interactions or relationships) that connect them and graph theories. SNA views social relationships (friendship, family, co-workers, etc.) in terms of network theory, also tries to study social networks to understand their behaviour and structure.

Analysis of social network data can be done by using the data mining based techniques that proved a significant impact especially for large datasets that cannot be handled by traditional techniques. SNA can work on any data source that contains interactions between entities (individuals or bank accounts.etc.) Moreover, it could be applied in different industries like:-

- Health

[Saldanha, et al., 2016]&[Poundstone, et al.,2004] utilized SNA in epidemiology to help understand how patterns of human contact inhibit or aid the spread of diseases such as Human Immune deficiency virus(HIV) in a population.

- Finance

SNA has an important role in the financial industry as detecting fraud cases. It also has an essential role in payments and banking. The social networks made up of card numbers, customers, accounts,

¹ Department of Computer Science, Institute of Statistical Studies and Research, Cairo University, Egypt
mohamed.mostafa.365@gmail.com

² Department of Information System, College of Computer and Information Sciences, Jouf University, Kingdom of Saudi Arabia, a.ahmed@cu.edu.eg

³ Vice-Dean of Institute of Statistical Studies &Research, Cairo University , Egypt , hehefny@ieee.org

phone numbers, email addresses and so on, and connections between these different nodes can be investigated to detect patterns that might indicate fraudulent behaviour.

- Telecommunication

In the telecommunication industry, customers profiles data used along with the Call Detailed Record (CDRs) generated from the Customer Service Provider to figure out the social relations between customers, in addition, combining this data with SNA will give a valuable insights that would have a great impact on the company profit and customer satisfaction. [Doyle, et .al.,2008] discussed why the telecom operators starting to see production deployments of SNA in the telecom sector is the emergence of technology designed to meet the specific needs of marketing. It focused on the relations between subscribers (customers), traditional propensity or segmentation models that are based on individual subscriber attributes. The relation between two subscribers can be identified using voice, calls, SMS, MMS, etc. The following are examples of how SNA is being used in the telecom sector across the world. The business applications include:

- Prediction of Churn and customer actions.
- Detecting rotational churn.
- Improving customer value management.
- Target the best candidate customer
- Improving up and cross-sell campaign performance.
- Influencer marketing.
- Fraud scoring.

- Security

[Krebs, et al.,2002] presented a study made by network scientist Valdis Krebs on the events of September 11, 2001, that investigate the terrorist network behind the attacks by piecing together connections between the hijackers from news reports and measuring three aspects of influence.

The rest of the paper is organized as follows, section II studied the different techniques of implementing SNA and their impact on social networks and finally, section III presents the conclusion.

2 SOCIAL NETWORK ANALYSIS TECHNIQUES

Many related works have been issued in SNA like detecting communities, influencers, customer acquisition and link prediction through the network.

Due to the big size of social network data, there is a need for a big data architecture to host the data and make the analysis on top of it. [Soric, et .al, 2017] discussed this issue as they proposed a new prototype platform of SNA upon the architecture of big data using Spark GraphX framework in combination with JavaScript.

2.1 Community Detection

Community detection is one of the most popular topics in SNA. Communities, or clusters, are groups of nodes having a higher probability of being linked to each other than to members of other groups. Many algorithms could be used to detect communities, the most frequent algorithm based on the fast unfolding technique and modularity of the nodes. It helps us to further get closely connected components as seen in figure 1. The algorithm is built on top of the community labelling. The link's weight is measured and used as input for this algorithm. The output of this algorithm is a membership vector that identifies the community ID to which every node belongs. [Altshuler, et al.,2012] took advantage of the accumulative nature of the mobile phone data, and they focused on building an incremental learning model for SNA to predict the personal properties like gender and nationality.[Stroele, et al.,2017] Applied SNA concepts to the Brazilian Scientific social network where each researcher in this network represents a node and their research represent edges between these nodes. They used centrality measures and clustering models to define the tie strength between researchers.[Cahyana, et al.,2017] proposed a solution for an issue in the SNA that can predict early behaviour that can be done by groups and actors, But it has shown weak groups and actors, due to using only one keyword, so the proposed solution used several different keywords, but have the same topic.[Zuo, et al.,2017] proposed a study for handling the centrality analysis by using uncertainty methods, they integrated multiple centrality measurements by using fuzzy set and MYCIN theory.[Pirim, et al.,2017] used mixed integer linear programming models to a dataset extracted from the social network. This research used two MILP clustering models M1 and M2. M1 model minimizes the maximum diameter between the cluster diameters; on the other hand, M2 reduces the total distance among objects related to the same cluster. M1 has computation time lower than M2.

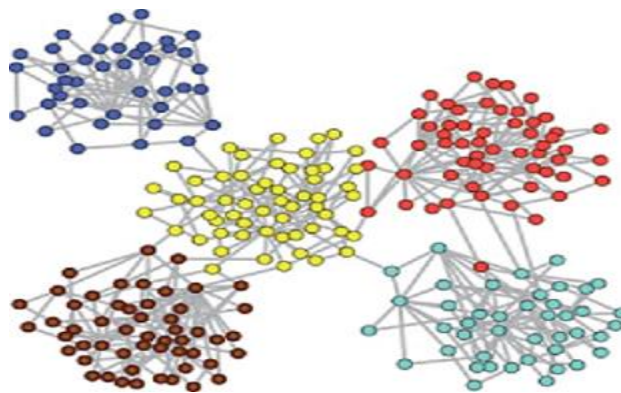


Figure 2-1 community detection

A comparison between the most common community detection algorithms was hold as shown in figure 2 in terms of three different evaluation metrics:

- Exec Time: Execution time with the incremental increase of number of nodes.
- NMI: The Normalized Mutual Information (NMI) has been widely used to evaluate the accuracy of community detection algorithms.
- Memory Usage: Memory consumption as the number of nodes increase.

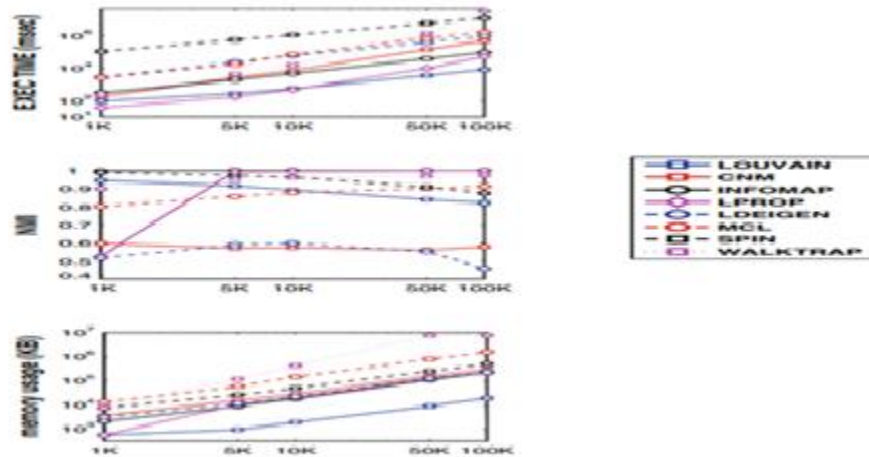


Figure 2-2 Benchmark results for common methods of Community Detection

2.2 Influencer Detection

An Influencer is the most connected node (customer) in the network. The detection of influencers is a recent line of research in SNA. In case of detecting and ranking influential people, two common approaches used widely in detecting key individual with a comparison as shown in Table 1:

- HITS Algorithm
- Page Rank Algorithm.

Table 2-1 Comparison between Hits & Page Rank Algorithm

Criteria	HITS	Page Rank
Basic Criteria	Link analysis algorithm	Link analysis algorithm based on random surfer model.
Main Technique followed	Web Structure Mining, Web Content Mining	Web Structure Mining
Efficiency	For a given a query HITS invokes traditional search engine to retrieve set of pages relevant to it and then attempts to find hubs and authorities. Since this computation is carried out at query time, it is not feasible for today's search engines, which need to handle millions of queries per day.	PageRank computes a single measure of quality for a page at crawl time. This measure is then combined with a traditional information retrieval score at query time. The advantage is much greater efficiency
Complexity Analysis	$O(kN^2)$	$O(n)$
Quality of Results obtained	Less than PageRank algorithm	Medium

The most common algorithm used to measure the score of each node is the page rank algorithm that is used to be applied on top of community detection phase. Each node in the network has a score that represents the influence of this node. The nodes with a high score are the most influential nodes. It has two main inputs, which are the graph and weights on the links in the graph. Several companies have produced tools, which supports marketing departments demand to detect leading users who make some influence on the followers and who will spread more effectively information about the products. This is necessary for the recent trend to invest in the so-called Word of Mouth marketing, where companies need to discover the influencers who will spread more effectively information about the new products.

Recent researches for SNA included mobile phone data, question answering community, mathematical programming, and how SNA can be implemented on big data platform, handling the uncertainty of centrality analysis, important actors analysis. For the mobile phone data that can give rich information on personal behaviour and their social relationships which are dynamic in nature. [Baruah, et. al., 2012] presented a study on the analysis of the related social network derived from mobile call detailed record. This research focused mainly on the evolution of the network, discover Influencers, find close associates of the influencers, and finally discovering customers who act with the same communication behaviour as the influencers. [Farooq, et. al. 2018] focused on visualizing and measuring the different metrics of social network and based on those values, it can detect the most influential node.

2.3 Customer Acquisition

Focusing on off-net customers are important to target them in campaigns, as they are actual potential customers to be acquired and included to your network. Accordingly, each node has been assigned an acquisition score calculated from the usage and ties along other on-net nodes. Further to that all off-net customers are categorized into (Low potential, Ideal, High potential). Node weight and some parameter as direction of tie and duration and strength and finally the regularity are used by scoring function to assign each node an acquisition score. The acquisition score is the result from using the scoring function, which categorized into two levels:

- Adding score on the link level between on-net and another off-net node.
- Adding Score on the node level between all on-net nodes and one off-net node that are accumulated from all scores of all links.

2.4 Link Prediction

A basic problem in SNA is the link prediction and the analysis of social profiles and their effect on anticipating the links in the social network. Therefore, [Eagle, et al. 2006] discovered that various kinds of social links performed different spatial-temporal proximity patterns. Focusing on a small similar population from the university. A prior work was done by [Resig, et al., 2004]& [Liben, et al., 2007]& [Fire, et al., 2011] for the purpose of predicting undiscovered and future links in a social network, However, the focus of their work is to characterize existing relations and predict their types. [Nanavati, et al., 2006]& [Onnela, et al., 2007]& [Seshadri, et al., 2008] used mobile call graph data to find and distinguish social interactions between subscribers of the cellphone. [Gonzalez, et al., 2008] studied how the mobility pattern changed with time in the urban areas. [Dong, et al., 2009]& [Onnela, et al., 2007] came up with different kinds of subscribers that behave differently; they investigated the strength of the links and their structure in cellphone networks. [Scellato, V., et al., 2010]& [Scellato, A., et al., 2011]& [Scellato, C et al., 2011] showed the importance of location and distance as key factors in making new links. [Cho, et al., 2011]& [Liben, et al., 2005] discussed the relation between mobility and friendship based on the physical distance. [Backstrom, et al., 2010] anticipated where the entities located in the social network. [Min, et al., 2013] discussed how we could use the communication patterns of the cellphone in identifying the relationship between the caller and callee. [Tang, et al., 2012] focused on finding social ties specifically in heterogeneous networks. [Wang, et al., 2011] used spatial-temporal features and network in anticipating connections in a large-scale cell phone network. [Crandall, et al., 2010] presented a methodology for predicting friendships by using spatial-temporal co-occurrence. [Jie, et al., 2015] discussed a suitable method for the dynamic multidimensional network to gain the actions that people want to see by analyzing the changes in the network. [Sharma, et al., 2017] carried on an experiment to evaluate the performance of baseline link prediction on simple and complex social networks. Three datasets have been taken and applied on them Adamic adar; Jaccard coefficient, preferential attachments and common neighbour. [Pan, et al., 2017] proposed a link prediction algorithm depended on structural perturbations of the weighted adjacency matrices. [Rahman, et al., 2017] came up with two different link prediction algorithms: Global, and local link prediction taking into account the activities of the users and the common friends. The similarity is measured between the two predicted nodes through two formulas called influential score based iscore and correlation based cscore. [Liu, et al., 2017] proposed a new

algorithm for anticipating the missing links in the network by using consensus clustering, assuming that grouping nodes together by community detection algorithms increases the similarity of them.[Xu, et al.,2017] studied and differentiated between the different similarity contribution of neighbour nodes and endpoints. In addition, substantiate the weak-link impact in networks.[Srilatha, et al.,2017] used swarm intelligence to detect structural similarity in the social network. They applied the firefly algorithm to predict links in the social network. [Linyi, et al.,2017] came up with an approach for predicting missing links, which is a collective index to fully quantify the common neighbour influence and inter-node influence to form the link prediction. The triadic closure structure is always quoted in the researches of network models as the fundamental structure of the local network and an essential link producing mechanism.[Qiu, et al.,2017] showed up two new indices, the first one was AD-SALP (Link Prediction based on the spectral Analysis adding Angular Distance) and the other one was SALP(Link Prediction based on the spectral analysis, using the classical Euclidean Distance, Manhattan Distance and Chebyshev Distance) to anticipate the missing links in the incomplete complex networks, that assigned weights to all edges in the network involving truly current and nonexistent links.

[Samad, et al.,2017] proposed a solution by evaluating the features and assign a new weight for those features.[Chen, et al.,2018] converted the link prediction problem into optimization problem (binary classification), where the Area under the Curve(AUC) is used to become an evaluation metric.[Chiu, et al.,2018] predicted links in social networks using deep learning. Proposing computationally efficient and network-size-independent feature vector with deep learning that is fit for the real-time application. [Xiao, et.al., 2018] used Twitter data for link prediction. Applying a three-level hidden Bayesian link prediction model by combining user behaviour and their relationships to find hidden links between users. All the previous studies made in this issue only discussed the evolution of the social network that is obtained from the call graph, but it did not provide any insights on how to use call graph data to characterize and classify the type of a social relationship between mobile phone subscribers using clustering algorithm like in figure 3.

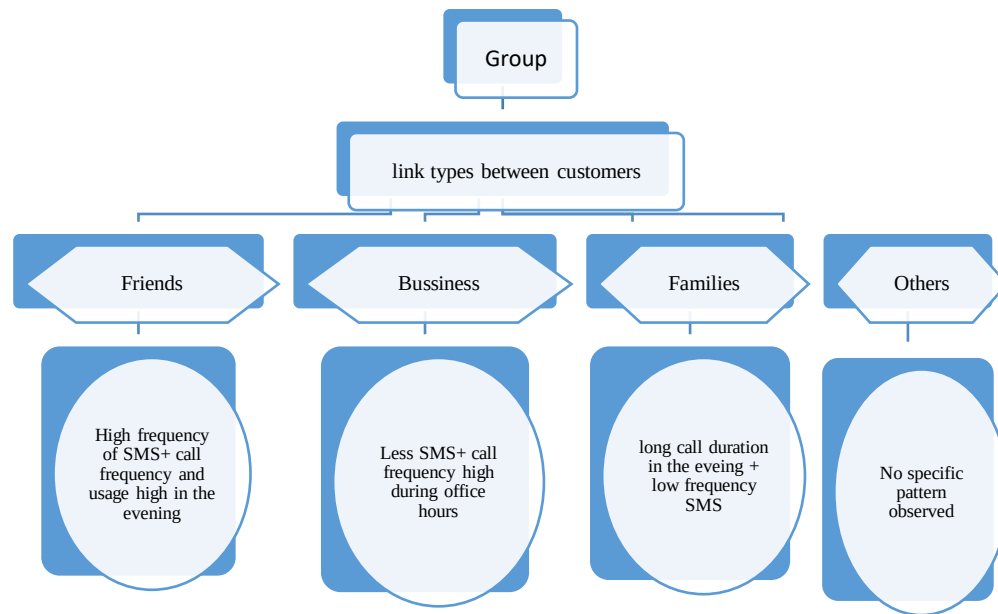


Figure 2-3 Flow chart for identifying the demographics of the group

3 CONCLUSION

Survey on Social Network Analysis revealed that SNA was recently applied in many different areas including Health, Finance, Telecommunications and Security. It could be used for detecting communities of customers, influencers who can attract others to be in the same community and acquiring customers in the social network. In addition, it raised the problem of predicting links that can be done in future between individuals through network and analyze the type of connection between nodes like (friendship, family, co-workers, etc.).

REFERENCES

Altshuler, N. Aharony, M. Fire, Y. Elovici, and A. Pentland,(2012).“Incremental learning with accuracy prediction of social and individual properties from mobile-phone data,” in Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). IEEE, pp. 969– 974.

Backstrom and C. Marlow,(2010). “find me if you can: improving geographical prediction with social and spatial proximity,,” in Proceedings of the 19th international conference on World wide web - WWW 10. New York, USA: ACM Press, pp. 61-70.

Baruah and P. Angelov,(2012).“Evolving social network analysis: A case study on mobile phone data,” in Evolving and Adaptive Intelligent Systems (EAIS), 2012 IEEE Conference on. IEEE, pp. 114–120.

Cahyana and R. Munir,(2017).“Community and important actors analysis with different keywords in social network,” in Science in Information Technology (ICSITech), 2017 3rd International Conference on. IEEE, , pp. 498–502.

Chen, Y. Hua, Y. Yuan, and Y. Jin, (2018). “Link prediction on directed networks based on auc optimization,” IEEE Access, vol. 6, pp. 122–136.

Chiu and J. Zhan, (2018). “Deep learning for link prediction in dynamic networks using weak estimators,” IEEE Access.

Cho and J. Leskovec, (2011). “friendship and mobility: user movement in location-based social networks,,” in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 11. New York, USA: ACM Press,, pp. 10-21.

Crandall, L. Backstrom and J. Kleinberg, (2010). “inferring social ties from geographic coincidences,,” Proceedings of the National Academy of Sciences of the United States of America, vol. 107, no. 52, pp. 22436-22441.

Dong, G.-J. Song and J.-Y. Wang, (2017). “an experimental study of large-scale mobile social network,,” in Proceedings of the 18th international conference on World wide web - WWW 09. New York, New York, USA: ACM Press, pp. 75-85.

Doyle, (2008). “Social network analysis in the telco sector marketing applications,” Journal of Database Marketing & Customer Strategy Management, vol. 15, no. 2, pp. 130–134.

Eagle and A. Pentland, (2006). “reality mining: sensing complex social systems,,” Personal Ubiquitous Comput, vol. 10, no. 4, pp. 52-68.

Farooq, G. J. Joyia, M. Uzair, and U. Akram, (2018). “Detection of influential nodes using social networks analysis based on network metrics,” in Computing, Mathematics and Engineering Technologies (iCoMET), 2018 International Conference on. IEEE, pp. 1–6.

Fire, M.; Tenenboim, (2011). “Link prediction in social networks using computationally efficient topological features,” in 'SocialCom/PASSAT', pp. 73–80.

Gonzalez and A.-L. Barabasi, (2008). “understanding individual human mobility patterns.,” Nature, vol. 453, no. 7196, pp. 77-82.

Jie, (2015). “Analysis of link prediction method in mobile social network,” in Measuring Technology and Mechatronics Automation (ICMTMA), 2015 Seventh International Conference on. IEEE, pp. 145–148.

Krebs, (2002). “Uncloaking terrorist networks,” First Monday, vol. 7, no.4.

Liben-Nowell and J. Kleinberg, (2007). “The link-prediction problem for social networks,” Journal of the American Society for Information Science and Technology, vol. 58, no. 7.

Liben-Nowell, J. Novak and A. Tomkins, (2015). “geographic routing in social networks.,” Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 33, pp. 11623-11628.

Linyi and L. Shugang, (2017). “The node influence for link prediction based on triadic closure structure,” in Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2017 IEEE 2nd Information. IEEE, pp. 761–766.

Liu, L. Yang, and S. Liu, (2017). “Improving the accuracy of link prediction using consensus clustering,” in Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2017 IEEE International Conference on. IEEE, pp. 1004–1010.

Min, J. Wiese and J. Zimmerman, (2013). “mining smartphone data to classify life-facets of social relationships.,” in Proceedings of the 2013 conference on Computer supported cooperative work - CSCW 13. New York, USA: ACM Press, pp. 285-294.

Nanavati, S. Gurumurthy and A. Joshi, (2006). “On the structural properties of massive telecom call graphs: findings and implications,” Proc. of 15th ACM CIKM, pp. 435–444.

Onnela, J. Saramaki and A.- L. Barabasi, (2007). “structure and tie strengths in mobile communication networks.,” Proceedings of the National Academy of Sciences of the United States of America, vol. 104, no. 18, pp. 7332-7336.

Pan, L. Gao, and J. Gao, (2017). “Link prediction in weighted networks via structural perturbations,” in Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2017 14th International Computer Conference on. IEEE, pp. 5–8.

Pirim, (2017). “Mathematical programming for social network analysis,” in Big Data (Big Data), 2017 IEEE International Conference on. IEEE, pp. 2085–2088.

Poundstone, S. A. Strathdee, and D. D. Celentano, (2004). “The socialepidemiology of human immunodeficiency virus/acquired immunodeficiency syndrome,” *Epidemiologic reviews*, vol. 26, no. 1, pp. 22–35.

Qiu, R. Hu, R. Yan, Y. Yao, and M. Jiao, (2017). “The new link prediction methods based on spectral analysis,” in *Semantics, Knowledge and Grids (SKG)*, 2017 13th International Conference on. IEEE, pp. 106– 112.

Rahman, L. R. Dey, S. Haider, M. A. Uddin, and M. Islam, (2017). “Link prediction by correlation on social network,” in *Computer and Information Technology (ICCIT)*, 2017 20th International Conference of. IEEE, pp. 1–6.

Resig, S. Dawara and A. Teredesai, (2004). “Extracting social networks from instant messaging populations,” In *Proc. of ACM SIGKDD*.

Saldanha, T. Li, C. Yang, C. Ugarte-Gil, G. W. Rutherford, and K. Dickersin, (2016). “Social network analysis identified central outcomes for core outcome sets using systematic reviews of hiv/aids,” *Journal of clinical epidemiology*, vol. 70, pp. 164–175.

Samad, M. A. Islam, M. A. Iqbal, M. Aleem, and J. U. Arshed, (2017). “Evaluation of features for social contact prediction,” in *Emerging Technologies (ICET)*, 2017 13th International Conference on. IEEE, pp. 1–6.

Scellato, A. Noulas and C. Mascolo, (2011). “socio-spatial properties of online location-based social networks,,” in *International AAAI Conference on Weblogs and Social Media*.

Scellato, C. Mascolo and V. Latora, (2010). “distance matters: geosocial metrics for online social networks,,” *Proceedings of WOSN*, vol. 10, p. 88-95.

Scellato and C. Mascolo(2011),“exploiting place features in link prediction on location-based social networks,,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 11*. New York, USA: ACM Press, p. 10461054.

Seshadri, S. Machiraju and J. Leskovec, (2008). “Mobile call graphs: Beyond power-law and lognormal distributions,” In *KDD*.

Sharma, S. K. Khatri, and L. Patnaik, (2017). “Performance evaluation of baseline link prediction techniques on simple and complex networks,” in *Electrical, Computer and Electronics (UPCON)*, 2017 4th IEEE Uttar Pradesh Section International Conference on. IEEE, pp. 573– 577.

Sorić, D. Dinjar, M. Štajcer, and D. Oreščanin, (2017). “Efficient social network analysis in big data architectures,” in Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2017 40th International Convention on. IEEE, pp. 1397–1400.

Srilatha and R. Manjula, (2017). “Structural similarity based link prediction in social networks using firefly algorithm,” in Smart Technologies For Smart Nation (SmartTechCon), 2017 International Conference On. IEE, pp. 560–564.

Ströele, F. Campos, J. M. N. David, R. Braga, A. Abdalla, P. I. Lancellotta, G. Zimbrão, and J. Souza, (2017) “Data abstraction and centrality measures to scientific social network analysis,” in Computer Supported Cooperative Work in Design (CSCWD), 2017 IEEE 21st International Conference on. IEEE, pp. 281–286.

Tang and J. Kleinberg, (2012). “inferring social ties across heterogenous networks,” in Proceedings of the fifth ACM international conference on Web search and data mining - WSDM 12. New York, New York, USA: ACM Press, pp. 743-752.

Wang, D. Pedreschi and A.-L. Barabasi, (2011). “human mobility, social ties, and link prediction,” in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 11. New York, USA: ACM Press,, pp. 1100-1108.

Xiao, X. Li, H. Wang, M. Xu, and Y. Liu, (2018). “3-hbp: A three-level hidden Bayesian link prediction model in social networks,” IEEE Transactions on Computational Social Systems, vol. 5, no. 2, pp. 430–443.

Xu and Y. Yin, (2017). “A similarity index algorithm for link prediction,” in Intelligent Systems and Knowledge Engineering (ISKE), 2017 12th International Conference on. IEEE, , pp. 1–6.

Zuo, B. Yang, and W. Zuo, (2017). “Exploring uncertainty methods for centrality analysis in social networks,” in Data Mining Workshops (ICDMW), 2017 IEEE International Conference on. IEEE, pp. 163–169.

A Deep Learning Approach for Text Generation

Belal Saeed Mahmoud¹

Mohamed Ezzeldin Saleh²

Ammar Mohammed³

Abstract

Many natural language processing applications use language models to generate text, these models are typically trained to predict the next word in a sequence, given the previous words. However, at test time the model is expected to generate the entire sequence from scratch. Automatic text generation is widely regarded as highly difficult problem, partially because of the lack of the data set; it tends to over-fit when the training dataset is small. In this paper, we introduce a corpus of Twilight and Alice in Wonderland stories as data sets. Based on the corpus, we will use bidirectional recurrent neural network (BRNN) for the text generation with the help of long short-term memory (LSTM) and we will perform two models one with word to vector (word2vec) and the other one with one-hot vector, we will use word2vec model with two datasets Twilight and Alice in Wonderland stories, and we will use the second model with Alice in Wonderland story only.

After fitting the previous models the results are:-

First model which with word to vector method ended with 0.00795, and with 0.9989 accuracy that's for Alice in Wonderland story, and for Twilight story \ended with .02716 loss, and with 0.9812 accuracy, second model which with one hot vector method ended with 0.13869 loss, and with 0.9608 accuracy.

Keywords: Text Generation, Deep learning, BRNN, LSTM, Word to vector, one-hot vector.

1. Introduction

Natural language processing is a subdivision of artificial intelligence concerned with automated interpretation and generation of human language. Wide varieties of applications depend on text generation, including machine translation, story generation, video/text summarization, question answering, among others. From a machine learning perspective, text generation is the problem of predicting

¹ Department of computer Science, Institute of Statistical Studies and Researches, Cairo University, Egypt
eng_belalsaeed@pg.cu.edu.eg

² Department of computer Science, Institute of Statistical Studies and Researches, Cairo University, Egypt
mohamed_ezzeldin@pg.cu.edu.eg

³ Department of computer Science, Institute of Statistical Studies and Researches, Cairo University, Egypt
ammar@cu.edu.eg

a syntactically and semantically correct sequence of consecutive words given some context. Popular choices for text generation models are language models based on n-grams, feed-forward neural networks, and recurrent neural networks. There are some definitions that we will introduce before we talk about the models that we will implement:

RNN is not like feed forward neural networks, RNNs can use their internal state (memory) to process sequences of inputs which helps in text generation. Bidirectional RNN is an RNN but the normal one is unidirectional one that learns only from one side but Bidirectional RNN learns from both sides which helps in learning text meanings and relation between words. Long short-term memory (LSTM) [Sundermeyer et al., (2012)] units are a building unit for layers of a recurrent neural network (RNN). RNN composed of LSTM units is often called an LSTM network. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell is responsible for remembering values over time intervals. One hot vector [Harris and Harris, (2007)] is a method for transforming each word in the dictionary to a vector of dictionary length of zeros except the word index in the dictionary takes one. Word2vec [Mikolov et al., (2013)] method which transform the word to several hundred dimensions with each unique word in the corpus being assigned a corresponding vector in the space, these dimensions are presenting the features of each word. In this paper, we propose a Bidirectional RNN with the help of LSTM method to generate new story based on model's learning from the data set. We will implement two word embedding methods; namely word2vector method, and one hot vector. Additionally, we implement the model and compare the performance of the two word embedding methods.

The rest of this paper is organized as follows; Section 2 gives background on fundamental concepts. Section 3 discusses some related work. Section 4 shows the proposed approach. Section 5 discusses the results. Finally, Section 6 conclude the paper.

2. Background

Natural language is the most natural form of communication for humans. It is therefore essential that interactive AI systems can generate text. Natural language processing (NLP) is an area of artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation. There

are several choices for text generation models such as language models based on n-grams, feed-forward neural networks, and recurrent neural networks. Feed Forward Neural Network [Tahmasebi and Hezarhani, (2011)] is the most basic artificial neural network that is used for regular regression and classification problem. The reason why it is called a feed forward neural network because there is no looping in the network. The information only flows in the forward direction in every layer of the network. The input layer accepts the inputs, feeds it to the hidden layers for all the calculations and manipulations. This calculated information is then forwarded to the output layer to produce the output as shown in figure 1.

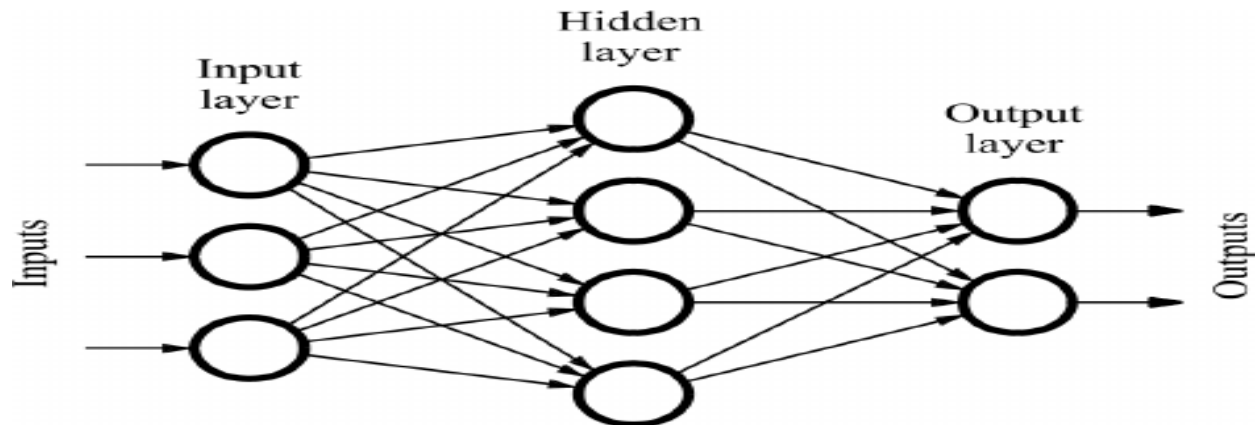


Figure 1 Feed-forward neural network [Quiza and Davim, (2011)]

Humans don't start their thinking from scratch every second. As you read this paper, you understand each word based on your understanding of previous words.

You don't throw everything away and start thinking from scratch again. However, Recurrent Neural Network [Ranzato et al., (2016)] works on the principle of saving the output of a layer and feeding this back to the input in order to predict the output of the layer. There is looping of information in every layer. RNN considers the current input and also the inputs received previously. Recurrent Neural Networks have the capability to memorize the inputs due to its internal memory. Each time step takes the input at the current time step as well as the input from the previous time step as shown in figure2.

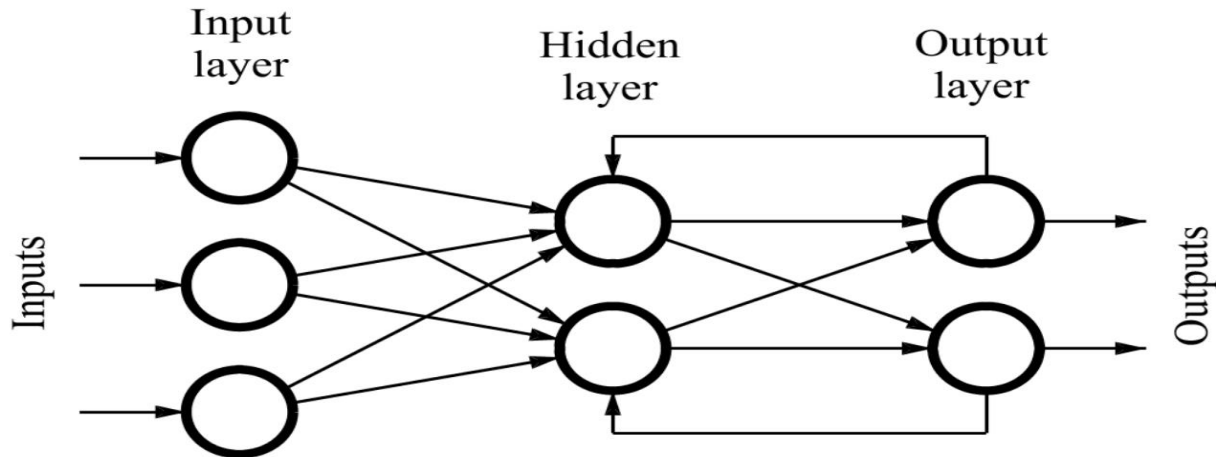


Figure 2 RNN [Quiza and Davim, (2011)]

RNNs are able to connect previous information to the present task, for example, consider a language model trying to predict the next word based on the previous ones. If we are trying to predict the last word in “Donald Trump is the president of,” we don’t need any further context – it’s pretty obvious the next word is going to be USA. In such cases, where the gap between the relevant information and the place that it’s needed is small, RNNs can learn to use the past information. RNNs also have restrictions as the future input information cannot be reached from the current state. However, BRNNs also have restrictions as the future input information is reachable from the current state. The basic idea of BRNNs is to connect two hidden layers of opposite directions to the same output. By this structure, the output layer can get information from past and future states. BRNN are especially useful when the context of the input is needed. For example, in handwriting recognition, the performance can be enhanced by knowledge of the letters located before and after the current letter. But there are also cases where we need more context. Consider trying to predict the last word in the text “I grew up in Italy I speak fluent Italian.” Recent information suggests that the next word is probably the name of a language, but if we want to narrow down which language, we need the context of Italy, from further back. It’s entirely possible for the gap between the relevant information and the point where it is needed to become very large. Unfortunately, as that gap grows, RNNs and BRNNs become unable to learn to connect the information. Long Short-Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies. LSTMs are explicitly designed to avoid the long-term dependency problem.

Remembering information for long periods of time is practically their default behavior, not something they struggle to learn. In natural language processing, a one-hot vector is a vector of N length where N is the vocabulary size, used to distinguish each word in a vocabulary from every other word in the vocabulary. The vector consists of 0s in all cells with the exception of a single one in a cell used uniquely to identify the word. Word2vec is a word embedding method, its input is a large corpus of text and produces a vector space, typically of several hundred dimensions with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space. T-distributed Stochastic Neighbor Embedding (t-sne) [Van der Maaten and Hinton, (2008)] is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low dimensional space of two or three dimensions.

3. Related work

[Jain, et al (2017)] discussed a model which takes sequences of short narrations and generates a story out of it. They have used phrased based Statistical Machine Translation to figure out target language phrase-maps and then merge all the phrase-maps to get the target language text. [Jaya and Uma, (2010)] proposed an ontology-based system where a user must provide characters, objects, and location to generate a story. Using ontology, attributes are provided to the selected parameter and story is generated.

4. Proposed Model

This section proposes an RNN model for automatic text generation. Also, this section describes the dataset that we will use to evaluate our model.

The proposed model is described as the following:

1. Create the dictionary:

We created the dictionary of the data set which contains the distinct words in the corpus.

2. Create sequence of words:

We created sequences of words with 30 words size each, which will be used in learning by learning the next word prediction according to the previous 30 words.

3. Creation of the neural network, we are using Keras that provides a very good abstraction to design a Neural Network architecture.

We created the following neural network:

I. Embedding layer: This layer can only be used as the first layer in a model, it turns positive integers (indexes) into dense vectors of fixed size, in our model its input is vocabulary size and its output is number of dimensions in word to vector, which is 300. This layer is used only in word 2 vector model.

II. Bidirectional LSTM layer: with size of 256 and using RELU as activation.

III. Dense layer of the size of the vocabulary, a dense layer is simply a layer where each unit or neuron is connected to each neuron in the next layer.

IV. Finally, a softmax activation layer

4. We use ADAM as optimizer and the loss calculation is done on the categorical cross-entropy. The network should provide us a probability for each word from the vocabulary to be the next one after a given sentence.

4.1 Dataset

The datasets are Alice in Wonderland [Carroll, (1865)] and Twilight [Meyer, (2005)] stories; Alice in Wonderland story contains 30318 words and vocabulary with 2984 words. This could be a small corpus, but we have chosen it because of data size constraint on using colab website, Twilight. Contains 630529 words and vocabulary with 15626 words, and we will use it only with word to vector method.

4.2 Data processing

The following preprocessing steps are performed to prepare the data for training:

1. Data cleaning:

- Transform all letters to small form to let the model knows that for Example: Sun is equal to sun because the model will deal with them as 2 words without this step.
- Replacing short form words to its original form for example replacing “I’m” to “I am”, but we have faced one problem for “’s ” form how to deal with it for example: John’s book ,and John’s good, “’s “ has different meanings in each one of them.
- Converting the corpus into list of words removing unnecessary characters, split the corpus to words only, no commas, no hyphens, and no full stops.

2. Transforming the words into a form that model can understand:

We have done two models each one is using a different method in this area, first model is using one-hot vector method which transform each word in the dictionary to a vector of dictionary length of zeros except the word index in the dictionary takes one, the second model used word2vec method which transform the word to several hundred dimensions (three hundreds in this model) with each unique word in the corpus being assigned a corresponding vector in the space, these dimensions are presenting the features of each word.

5. Experiment result:

The results are shown in Table 1.

Table Head	Models type		
	Text generation with w2v (Wonderland dataset)	Text generation with w2v (Twilight dataset)	Text generation with one hot vector
Start loss Fn	7.4388	8.7588	6.7612
End loss Fn	0.0079	0.02716	0.13869
Start accuracy	0.0331	0.03541	0.0348
End accuracy	0.9989	0.9812	0.9608
Epochs	120	80	160

Table 1: Comparison between results

Even that Word2vec model with Wonderland dataset has the best results according to the accuracy and loss function, it also has the worst meaningful output, that's because of the small amount of data so the model can't learn the similarity between words, we used t-sne method to prove this point for example the most similar words to "boy" are (lake, stolen, image, yawing, melancholy, met) with similarity between (.16 to .18) which is very bad as shown in figure 3.

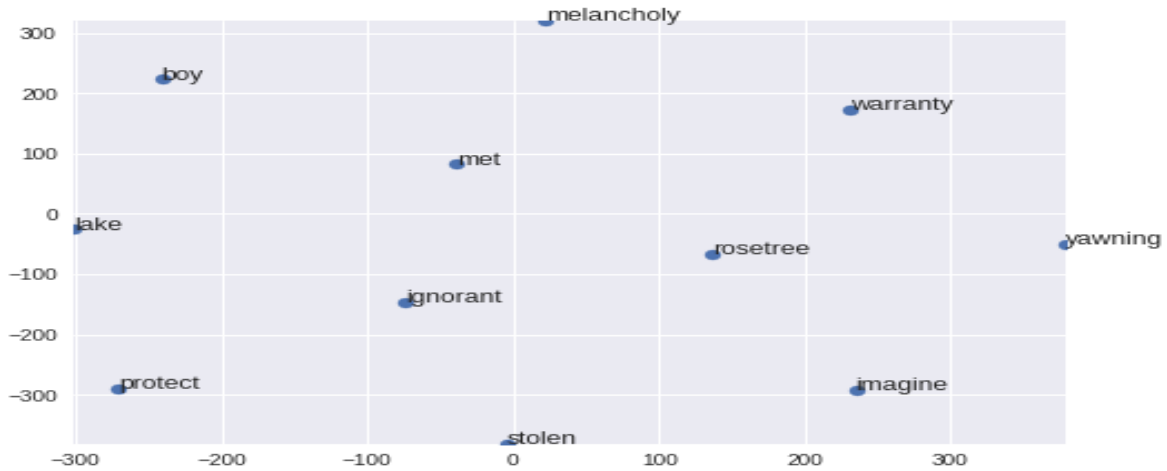


Figure 3 t-sne representation of word “boy”

We can comparison by loss function between word 2 vector and one-hot vector shown in figure 4 and 5.

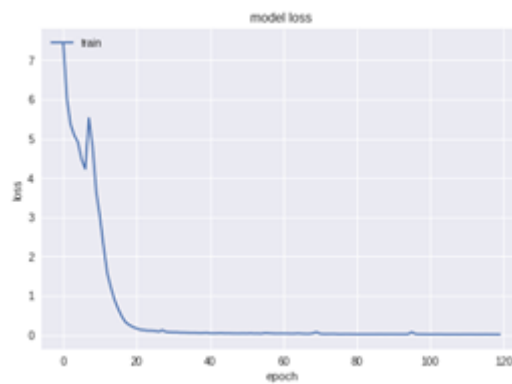


Figure 4 word2vec model loss function

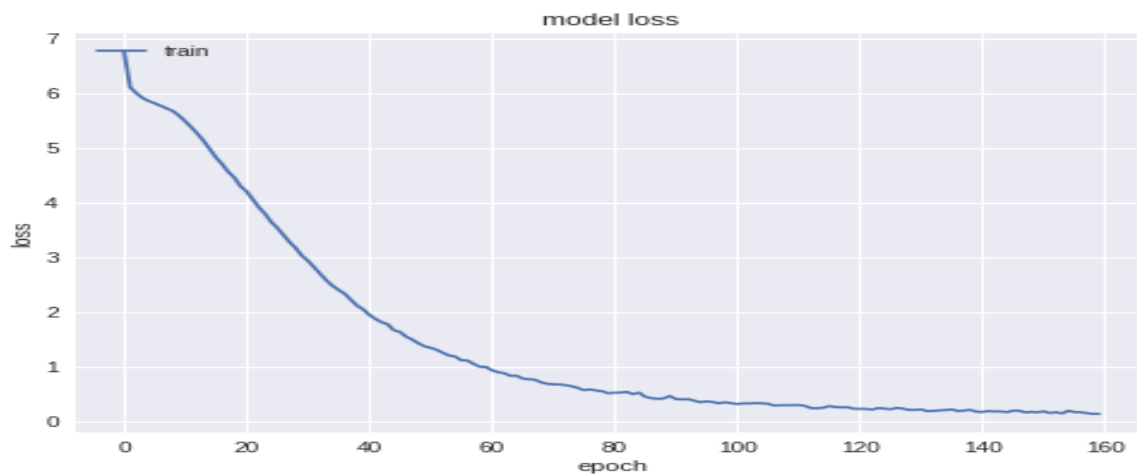


Figure 5 4 one-hot vector model loss function

We can comparison by accuracy between word2vec and one-hot vector shown in figure 6 and 7.

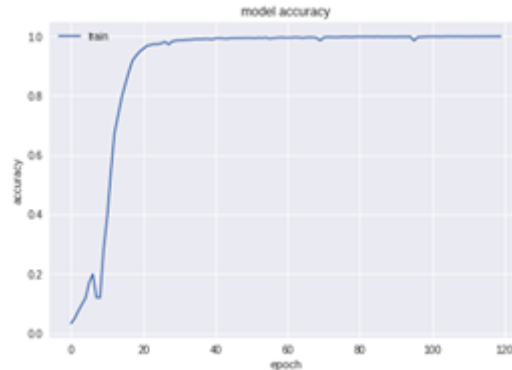


Figure 6 word2vec model accuracy

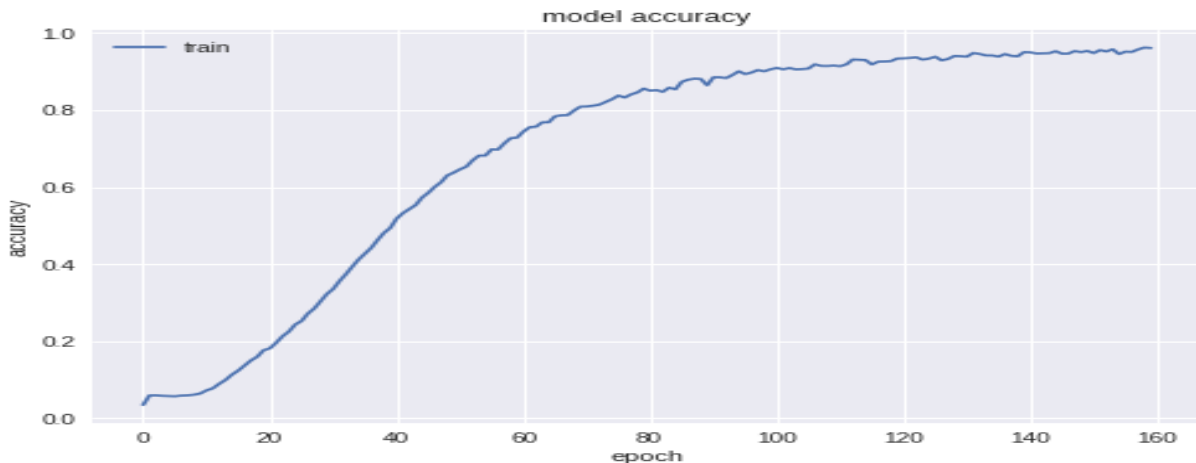


Figure 6 word2vec model accuracy

The input for them was: “she took down a jar from one of the shelves as she passed”
 The output for word to vector was: “yelp australia unfortunate ix butterfly today facility is note howling fireplace ornamented brandy puss uglification considerable royalty flappers wonderland breach give blew magpie poured checked despite be teacups walks repeating bowed uncommonly character bells head answers leaving eating liked org subjects hurrying treated refund matter conger memory couples site row”.

The output for one-hot vector was:” and is is the lives she was close and as off when she saw the whole little little sharp key so she saw the back of the whole pack at crowded at one round she had read out at the cook of time when the baby took out it”. One-hot vector got worse results than word 2 vector in loss function and accuracy but got more meaningful result that word to vector, that’s because it treats each word as unique word and because of the small amount of dataset in this story it got better

results as output. Wordtovec model with Twilight story got more meaningful results than the outputs for both models with wonderland story, because of the large amount of corpus for Twilight story.

6. Conclusion

NLP applications is the most effective way to allow computers to communicate with humans. In our experiment we used language models to generate story, these models are typically trained to predict the next word in a sequence, given the previous words. However, at test time the model is expected to generate the entire sequence from in this paper, we used a corpus of Wonderland and Twilight stories as data set. Based on the corpus, we used bidirectional recurrent neural network (BRNN) for the text generation with the help of long short-term memory (LSTM) and we performed two models one with word to vector (word2vec) and the other one with one hot vector. After the models was fitted the models we generated some text based on the stories, the final results with word to vector with Wonderland story was the best results according to accuracy and loss function, however it was the most unreadable results that's because of the small size of this story so it tended to over-fit with word2vec method, the results with Twilight dataset was much better because of the large amount of the data set, the second model generated a better text for Wonderland story because it treats every word as unique one so no fake similarity came from the small dataset.

References

Carroll, L., (1865), Alice in Wonderland. Macmillan, London.

Harris, D. and Harris, S., (2007). Digital Design and Computer Architecture. Morgan Kaufmann Publishers, Massachusetts.

Jain, P., et al (2017). Story Generation from Sequence of Independent Short Descriptions. In Conference on Knowledge Discovery and Data Mining.

Jaya, A., and Uma, G.V., (2010). An intelligent system for semi-automatic story generation for kids using ontology. Proceedings of the third Annual Computer Conference, Bangalore Conference.

Meyer, S., (2005), Twilight series. Little, Brown and Company, Columbus, Georgia.

Mikolov, T., et al., (2013). Distributed Representations of Words and Phrases and their Compositionality. In Advances on Neural Information Processing Systems.

Quiza, R. and Davim, J. P., (2011). Machining of hard materials. Springer, New York.

Ranzato, M., et al., (2016). Sequence level training with recurrent neural networks. In International Conference on Learning Representations.

Sundermeyer, M., et al., (2012). LSTM neural networks for language modeling. In INTERSPEECH.

Tahmasebi, P. and Hezarhane, A., (2011). Application of a Modular Feedforward Neural Network for Grade Estimation. Neural Resources Research.

Van der Maaten, L. and Hinton, G. E., (2008). Visualizing data using t-SNE. Journal of Machine learning Research.

AFSCloud Uniform Restful API Data Access for Cloud Data Store

Sherif Younis Saleh¹ Fatma El Licy² Akram Salah³

Abstract The world of technology had witnessed a huge evolution in the software deployment platforms of cloud computing. As NoSQL data management is used to generate client data, cloud computing has a great touchable effort in managing large volumes of this generated data. NoSQL applications vary according to clients' requirements of scalability, fault tolerance, and availability. As there is no standard model provided by cloud providers, cloud interoperability and portability become a major challenge for facilitating adopting cloud environments.

This paper proposes a common Restful API for cloud-based to solve the battle of data interoperability and portability between different cloud-based data storage services such as, Amazon RDS, Google Datastore, Amazon SimpleDB, couchDB, MongoDB, Blob storage, and Queue storage.

Finally, a case study is proposed to prove the implementation of the framework. The YCSB Benchmark is used for the system evaluation.

Keywords: Uniform data access, Relational database management systems, NoSQL database management, interoperability, Vendor lock-in, Multi-cloud, YCSB and storage systems.

1. Introduction

According to recent literature studies (Brian F et al., 2010), cloud interoperability is considered to be one of the most popular reasons for customers to be afraid of adopting cloud computing (Brian F , et al, 2010). Interoperability is defined to be the possibility of operating and dealing together for different heterogeneous applications. While as a cloud, interoperability term is used to describe the ability for different cloud vendors to handle the process of acting together, determining application standards, understanding Service Level Agreement (SLA) templates, authentication and authorization token formats and attribute data of each other's (Piyush Harsh et al., 2012).

Clients could save data in a cloud using the cloud computing property of storage-as-a-service A. (Rafique et al., 2017). Commonly among cloud providers, there are four categories of cloud data storage and data access mechanisms (Livenson I et al., 2011):

- Blob storage approach to store unstructured or semi-structured data, Examples includes *Amazon S3* and *IBM Cloud Object Storage*.

¹Department of Computer Science, Institute of Studies and Researches, Cairo University, Egypt
Sherif_bn_3m@yahoo.com

²Department of Computer Science, Institute of Studies and Researches, Cairo University, Egypt

³Department of Computer Science, Institute of Studies and Researches, Cairo University, Egypt

- The "NoSQL" mechanism, which is responsible for storing a table in a non-relational database. Examples include *Amazon* and *DynamoDB*.
- Queue storage approach for the totally handled queuing applications, examples include *Amazon* and *SQS*.
- Relational table storage for traditional databases, For example, *Amazon RDS*, and *Azure SQL databases*.

One of the most complicated problems is the problem of moving data or software from one cloud provider to another. The cost needed for rebuilding an application to coop with new cloud provider is called the migration cost. This cost is calculated based on, not only, the cost of developing a new system, but also the time and cost of educating developers the new provider's proprietary API to manage data storage.

A proposed unified Restful API (AFSCloud) is presented and founded in this paper to face these documented challenges. The main motive is controlling the stored data regardless of its storage format - that are supported by different cloud providers. AFSCloud introduces standards for restful APIs that could be adopted for Blob storage, NoSQL, Queue storage, and traditional database. AFSCloud is designed to give cloud users and developers the chance to free themselves from the cloud providers' restrictions, as their data would not dependent on cloud providers anymore. The proposed AFSCloud supports cloud developers to manage and move their data freely among cloud providers, as follows:

- For Blob storage, it supports CDMI Blob and non CDMI Blob,
- For *NoSQL* it supports *Google Datastore*, *Amazon SimpleDB*, *couchDB* and *MongoDB* systems,
- For the message queuing service, it supports CDMI Queue and non CDMI Queue.
- For traditional Database system, it supports *Amazon* provider namely, *AmazonRDS*.

Table 1, illustrates the relations between concepts terminologies in the respective data model and the proposed concepts terminology in order to define the AFSCloud.

Table1: Concepts Comparison

	AmazonRDS	Mongo DB	Google DataStore	SimpleDB	Couchbase	AFSCloud
Data model	Database	Document	Key-value	Column family	Document	
group	Table	Collection	Kind Name	Domain	Bucket	Type
Support Relation	yes	yes	No	No	No	yes
query	sql	Custom API	GQL	Custom SQL	N1ql Query	sql
operation	Custom API	Custom API	Custom API	Custom API	Custom API	Custom API

The rest of this paper is organized as follows: section 2 contains the literature discussion, section 3 explain, in details, the system framework of AFSCloud, section 4 discusses The AFSCloud API, and the services it provides, Section 5, presents a case study for the evaluation of the proposed framework. Finally, section 6 presents the conclusion and future work.

2. Related Work

Studying the literature makes it easy to notice that some of the researchers had already gave attention to different types of cloud data storage. Spring Data Framework (Pollack M et al., 2011). presents some comprehensible themes to direct the main categories of *NoSQL* DBMSs and relational DBMSs. These elementary studies are smoothed for each DBMS. The model provided by (Pollack et al., 2011) is basically based on a concordant developing approach using a set of programming samples and abstractions determined by the Spring Framework. The solution approach provided by them is deeply related to the Java programming model.

In a modern attempt to apply a promising solution, SQL++ (Ong K. W et al., 2011) fulfills an uninhabitable semi-structured data model which could be added to the relational data model, and that being developed using the JavaScript Object Notation (JSON).

JSON is used to express the relational tables; they (Ong K. W et al., 2013) worked basically on the mapping stage to present each relational tuple in a corresponding JSON literal. The model called SQL++ because it contributes in the SQL++ query language, as all data stored in this model could be queried and manipulated with the SQL++ query language. They claimed that their model is able to handle any kind of DBMSs.

NoSQL data models are also had a great effort in the literature. Object-*NoSQL* Datastore Mapper (ONDM) (Cabibbo, 2015) is a temple structure that purpose to help storing objects and retrieving them in *NoSQL* data stores. Although it prioritizes *NoSQL*-based applications developers with an Object Relational Mapping-like (ORM-like) API (e.g. JPA API), ONDM does not paid enough attention and effort to the relational data stores that would face *NoSQL* developers.

In other hand, to initiate an approach to optimize using benefits from several cloud providers, (Bastiao et al., 2013) present a Service Delivery Cloud Platform (SDCP). The SDCP is a novel infrastructure that could be best described as a cloud middleware. It catches resources from multiple cloud providers to give a group of services. They claimed to enhance the SDCP-SDK to develop SDCP-based applications. Developing an API is counted like a main contribution for (Bastiao et al., 2013). This API is made to enhance the cloud services of storage, database and notification service. Their API has disparate characteristics that would be applied in several cloud providers. The authors claimed that, these API support, also, encryption and decryption layers that were implemented on the client side.

Enhancing the user-cloud provider dependency, SimpleCloud (Vijaya et al., 2015) is launched. It is an API to assure the idea of allowing the usage of the storage services independent of cloud platforms. Allowing developers to code and interoperate with multiple cloud vendors, the SimpleCloud, supports, only, PHP language for web applications. This violates the fact that a cloud provides a joint interface to handle and manipulate the data stored in the multiple data centers upon different cloud providers, regardless of the programming language used by developers.

As for *NoSQL* databases, CDPort (Sakr et al., 2014) introduced an integrated data model and a standardized API. Providing only *NoSQL* storage type, CDPort supports the transformation and exchange of data that is stored on *NoSQL* databases.

The CDMI (Cloud Data Management Interface)(SNIA, 2017), had been introduced by the SNIA (Storage Network Industry Association). CDMI is an ISO/IEC standard that was proposed to solve the problem of the growing need of interoperability for data stored in the cloud faced by cloud solution vendors. Standards, provided by CDMI, face a clear shortage of acceptance and participation by Open Cloud Computing cloud providers.

A novel data integration methodology appeared on the area of cloud interoperability introduced by (Vathy-Fogarassy et al., 2017). Their target was facilitating the way of querying data individually from different relational and NoSQL databases. The approach provided collects data from separated database management systems according to the filtering options and migrates them. But on a deeper study of their proposed solution, it is so clear that it does not support joins and aggregates across data sources. They depend basically on a metamodel approach and it covers the structural, semantic and syntactic heterogeneities of source systems. a web-based application was introduced to apply their methodology practically. This application succeeded to confirm the usefulness of the suggested method.

Literature never lack programming solutions, for resolving the cloud interoperability problems. To easily access NoSQL and relational data stores referred to as Save Our Systems (SOS), (Atzeni et al., 2012) .introduced a programming interface. SOS is supposed to be an intermediate database access phase between an application and the different data stores that it uses. To do so, the authors defined a common interface to access different NoSQL DBMSs and a common data model to map application requests to the target data store. They argue that SOS can be extended to integrate relational data store. Studying this solution and analyzing the corresponding results never led to proof of the efficiency and the extensibility of their system.

Table 2 is a summary of the limitations of the discussed research papers regarding cloud interoperability. These limitations might be resolved by the proposed framework.

Table 2: Summary of the limitation of current cloud interoperability approaches

Related work	Limitations
Spring Data Framework	Language dependent.
SQL ++	Supports, only relational data model and JavaScript Object Notation (document database).
ONDM	Language dependent and do not support relational data stores.
SDCP	Language dependent.
SimpleCloud	Language dependent.
CDPort	Support NoSQL only
CDMI	Do not support NoSQL storage and relational table storage.
Vathy-Fogarassy et al	Support data integration only.
Atzeni et al	no proof of the efficiency.

3. AFSCloud Framework

The Architecture of the proposed framework for AFSCloud will be introduced, its dedicated data model will be explained and the present the AFSCloud API, will be presented, in the following subsections.

3.1 Architecture of AFSCloud

The main goal of the proposed framework is creating standardized Restful API to allow cloud developers to be independent from cloud vendors. Such API, will help in migrating data without re-engineering their associate applications.

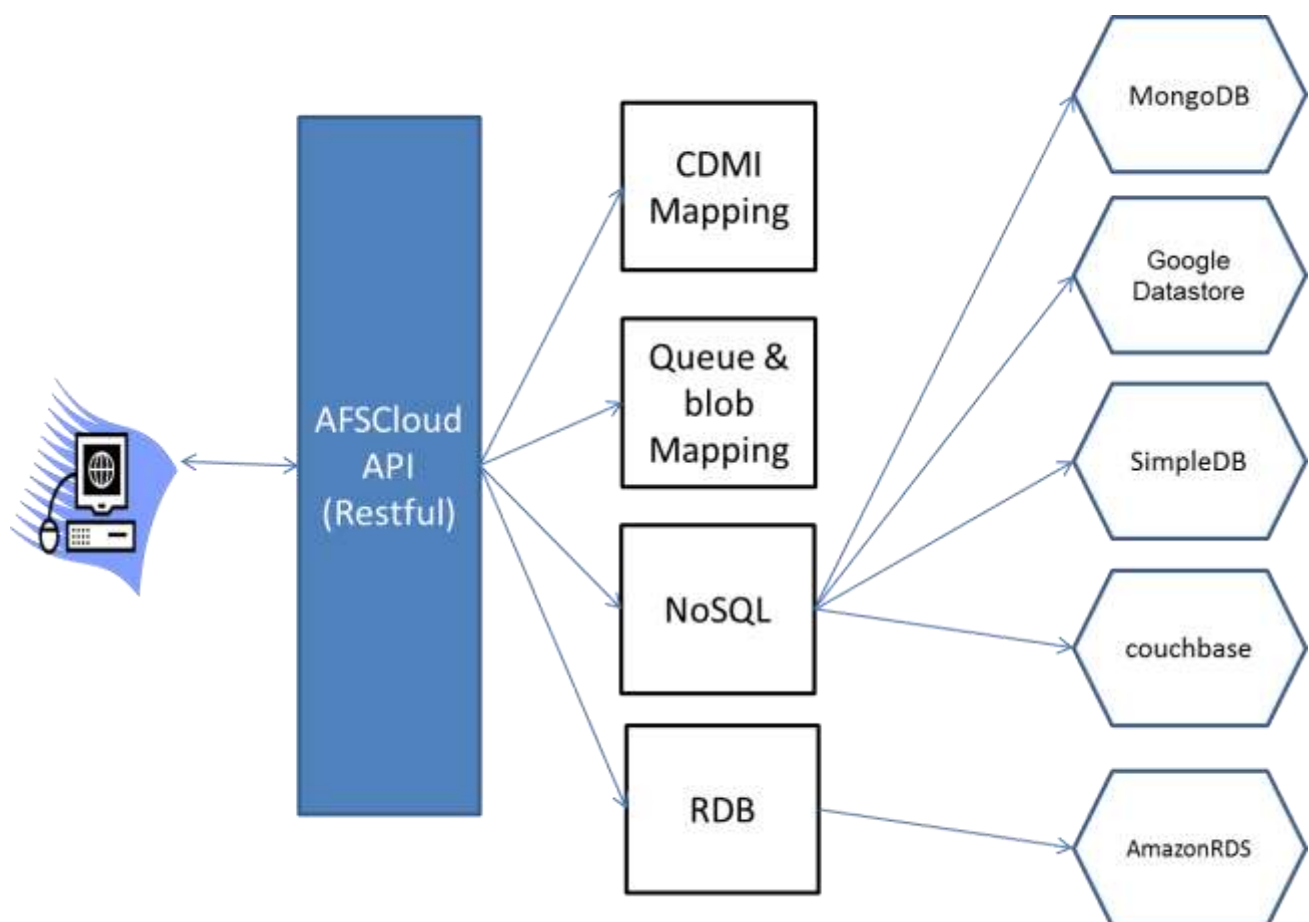


Figure 1: Architecture of AFSCloud framework

Figure 1, illustrates the architecture of the proposed framework. It consists of four components, namely: client/source, API, mapping layer and the destination API's follows:

1. Client: Applications that uses AFSCloud API.
2. AFSCloud Restful API: carries out the CURD and data transform operations.

3. The mapping layer: maps AFSCloud returned entities to specific cloud vendor API. This layer will solve conversions issues as follows:
 - a. Key type: the data type for each key differs according to the data models. Example: SimpleDB supports string type keys only while MongoDB enables using string, long or ObjectID type keys. Google Datastore supports string or long type keys.
 - b. Data type: to handle the issue of different data types, the proposed framework makes transparent conversion to eliminate the manual conversion for data types, when changing the backend storage. To illustrate, let us assume an application that interacts with MongoDB as the backend and store entities with different data types. As a result of this, changing the application to interface with another datastore such as SimpleDB would require the developer to convert all data types to String as this is the only data type supported by SimpleDB.
 - c. Standard query: a unified query abstraction interface is provided by the framework to enable the client to utilize core constructs of the SQL query language (e.g., SELECT, INSERT, DELETE, UPDATE).
4. The APIs of cloud vendors that are supported by AFSCloud.

3.2 AFSCloud Data Model

Figure 2, illustrates the proposed AFSCloud data model. It is all about unifying the principles of dealing with different NoSQL models to produce a standardized interface.

In the proposed data model, data objects are stored as entities where each entity stores the following information:

- Entity: It is a mandatory categorization attribute for the entity.
- Key: It represents an identification attribute for the entity which must be unique through the entity type. This key attribute can be of data type string, integer or object ID so as to be compatible with the different NoSQL systems.
- Attribute: It is a descriptive key-value attribute that store the attributes of the entity. An entity may be string, long, float, date, Boolean or an array data value. Apparently, each entity can have multiple descriptive properties.
- Entity Parents: are references from one entity type to another. Using the value of the first entity field (_id), collection name represents relations between entities that would be used to learn ontology for data integration in the future version.

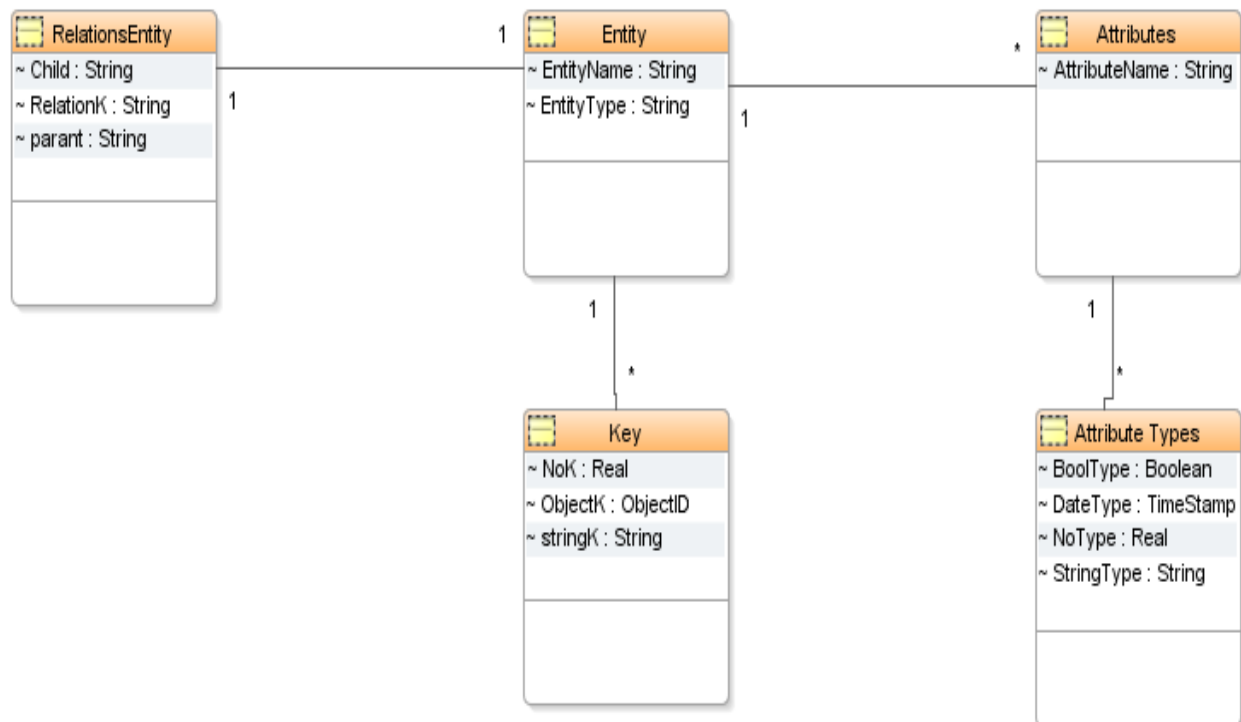


Figure 2: AFSCloud Framework Data Model

3.3 AFSCloud API

AFSCloud API constitute four kinds of operations. The first options family is dedicated to get CDMI operations. The second, operations are dedicated to Blob and queue. The Third operations are dedicated to traditional database and *NOSQL*. The fourth operations are dedicated to transformation between traditional database and *NOSQL* types. In particular, AFSCloud API supports the following methods:

- Put: This method is used for adding a new entity with its identification information and name. The method uses a Boolean parameter, which indicates that the entity should be replaced if it exists when the value of the parameter is set to true.
- getEntity: This method is used for retrieving a specific or all entities that belong to the same specified entity type.
- getProperty: It is used for retrieving a specific property or all properties for a specified entity.
- deleteEntity: It helps to remove an existing specific or all entities of a specific entity type.
- deleteEntityType: The method supports deleting the entity type and all its associated entities.
- listofEntityTypes: This method offers retrieving a list of entity types.
- Query: This method receives the query statement and returns the result back to user.
- setParent: This method uses to set parent document id similar to relationship in relational database.
- getParent: This method retrieving parent document id.

3.4 Data Exchange

AFSCloud support transformation data between any two data modules. Figure 3, explains AFSCloud data exchange steps as follows:

1. Create connection to source and destination data providers.
2. Build query to get data from source.
3. Retrieve data from source and store it in AFSCloud metadata file.
4. Convert data two AFSCloud data model.
5. Insert data to destination providers.
6. Optionally, remove data from source after insert in distinction.

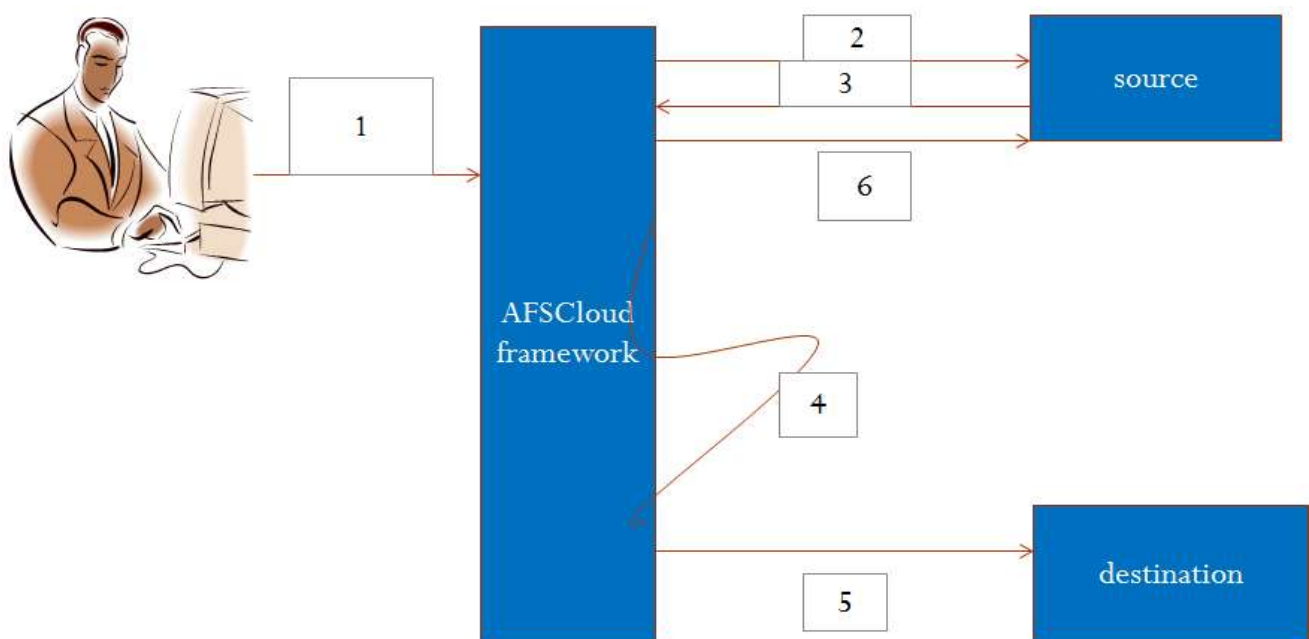


Figure 3: AFSCloud data exchange steps.

4. Implementation

AFSCloud API has been designed with the goal of hiding the programmatic differences between different clouds. Also, data storage services are supported by the framework to provide a unified programming interface. The AFSCloud API has been built using *java restful* web services.

5. Evaluation

The proposed framework is devised as a prototype. It is to be evaluated with respect to:

- Practicality and ease-of-use of the AFSCloud standard data model and API.
- Efficiency of data Transfer between different data model.

To simulate the case of developing portable java application on different *NoSQL* systems, the YCSB workload generator was utilized to generate a workload (WL) with simple data set and four

operations. The operations in the generated workload represented a balanced mix of read, insert, scan and update operations. We have uploaded the generated workload into each of the five supported databases (SimpleDB, MongoDB, Google Datastore, couchDB and Amazon RDS). Then, we represented the YCSB's generated workload using the methods of our proposed framework API.

In particular, we started by executing the generated workload over the SimpleDB database using the SimpleDB Mapping of the framework. On the second step, we assumed that we need to change the backend of our workload to the MongoDB database, therefore, we have changed the used Mapping to the MongoDB Mapping with the required credential and configuration information (the connect method). We have noticed that this was the only change we need to make in our code and the workload can be successfully executed over the MongoDB backend with no errors. Finally, we followed the same scenario with the Google Datastore backend and similarly we only needed to change the adapter class to create an instance of the new Database that we need to use as our backend without any need to make other changes in the rest of the application's code and also without the need to take any consideration regarding the difference between data models of the underlying database.

In the following subsections we present and analyze the execution times based on reads and updates operations at the same time. We executed YCSB workloads A.

Figure 4 shows the results, in seconds, obtained while executing workload A that consists of 50% reads and 50% updates, over 600.000 records and was run over a 16GB RAM, 1.86GHZ Intel Core7i machine . The execute workload A shows that update and read is slower with AFSCloud than other common API CDport (java API) and original support cloud data management API .

To simulate the case of Transfer data between different data model, we have used the YCSB workload generator to generate a workload with simple data set then upload data set SimpleDB, then transfer using client java Application as interface to restful transformation service as shown in figure 5 as following:

- SimpleDB → MongoDB.
- MongoDB → Google Datastore,
- Google Datastore → couchDB.
- couchDB → Amazon RDS.

The outcome of this case study shows the practicality and ease-of-use of the AFSCloud standard data model and API.

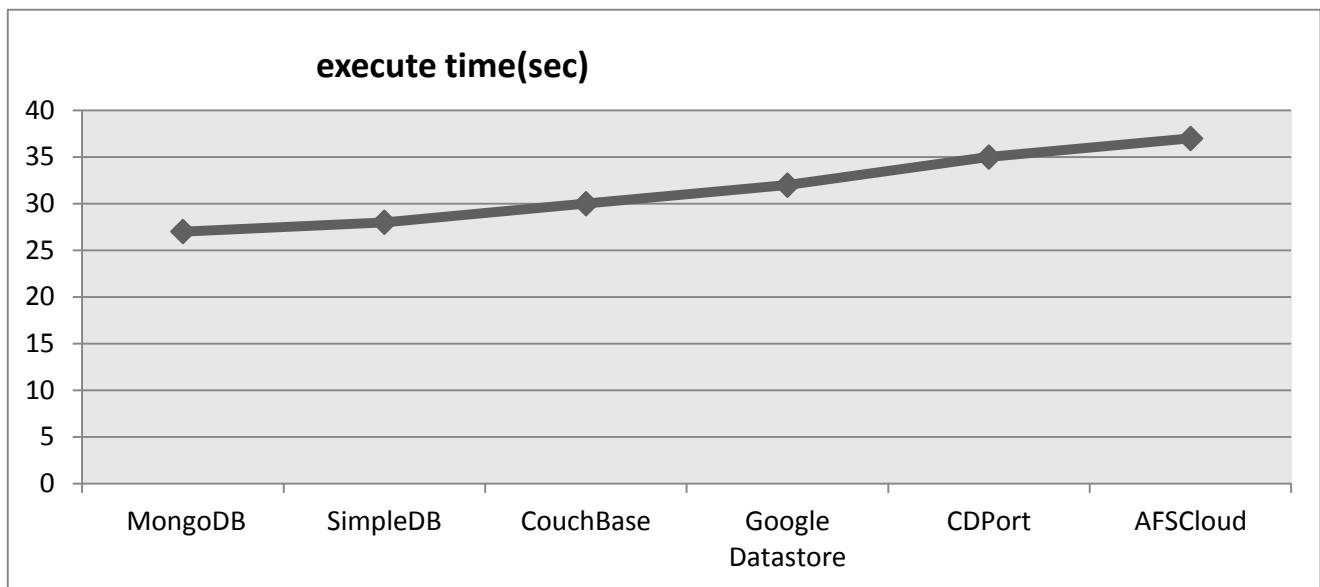


Figure 4: Execution time of workload A

6. Conclusions

In this paper we introduced a Restful API that enables developers to build their applications with independent cloud vendors API. AFSCloud provided abstract layer that support the CRUD operations and querying of NoSQL databass defined in Google Datastore, Amazon SimpleDB, couchDB, MongoDB, Amazon RDS using a common set of operations API.

In addition, AFSCloud support service to transfer and exchange date between different data model which may be divided into three kinds of transferring. First, transfer between nosql and relational database. Second, transfer between queues. Third, transfer between blobs data storage.

Finally, we describe the design and the proof-of-concept implementation of our framework using a case study which evaluated with Yahoo! Cloud Service Benchmark (YCSB).

As a future work, we plan to extend the implementation of our framework to supporting other NoSQL complex data models such as the graph model (e.g. Neo4J) and supporting semantic data integration using ontology.

7. References

Brian F. Cooper, Adam Silberstein, Erwin Tam, Raghu Ramakrishnan, and Russell Sears , (2010). "Benchmarking cloud serving systems with YCSB". In SoCC, pages 143-154.

Piyush Harsh, Florian Dudouet, Roberto G. Cascella, Yvon J_egou, and Christine Morin, (2012). "Using open standards for interoperability issues, solutions, and challenges facing cloud computing". In CNSM, pages 435-440.

Rafique A, D. Van Landuyt, V. Reniers, and W. Joosen, (2017). "Towards an adaptive middleware for efficient multi-cloud data storage".

Livenson I and E. Laure, 2011. "Towards Transparent Integration of Heterogeneous Cloud Storage Platforms".

Pollack M, O. Gierke, T. Risberg, J. Brisbin, and M. Hunger, Eds., Spring Data , 2012. "O'Reilly Media", October.

Ong K. W, Y. Papakonstantinou, R. Vernoux, (2013). "The SQL++ semistructured data model and query language: A capabilities survey of sqlon- hadoop, nosql and newsq databases", CoRR abs/1405.3631.

Cabibbo L, 2015. "Ondm: an object-nosql datastore mapper," Faculty of Engineering, Roma Tre University. Retrieved June 15th.

Luis A. Bastião Silva, Carlos Costa, and Josie Luis Oliveira, (2013). "A common API for delivering services over multi-vendor cloud resources. Journal of Systems and Software", 86(9):2309{2317.

Vijaya A and V. Neelanarayanan, (2015). "Framework for Platform Agnostic Enterprise Application Development Supporting Multiple Clouds" Procedia Comput. Sci., vol. 50, pp. 73–80.

Alomari E. A. Barnawi, and S. Sakr, (2014). "Cdport: A framework of data portability in cloud platforms," in Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services, pp. 126–133.

"Cloud Storage Initiative | SNIA." [Online]. Available: <https://www.snia.org/forums/csi>. [Accessed: 05-Jun-2017].

Vathy-Fogarassy, Tamas Hugya, (2017). "Uniform data access platform for SQL and NoSQL database systems, Information Systems".

Atzeni P, F. Bugiotti, and L. Rossi, (2012). "Uniform access to nonrelational database systems: The sos platform," in Advanced Information Systems Engineering - 24th International Conference, CAiSE 2012, Gdansk, Poland, June 25-29, 2012. Proceedings, pp. 160–174.

A Survey of Quantum data compression, Methods, and Approaches

Mohamed Nehad¹, Fatma A. EL-Licy¹, Akram Salah²

Abstract

Quantum data compression is an essential subject for the researches of Quantum computing, physics, and mathematics. There are a lot of methods and approaches raised up in the last 2 decades, some of them uses the mathematical modeling to compress the quantum data, some use the physics concepts to engineer or enhance the devices, others use the neural networks in the compression process, and others enhances the existing classical data compression to apply it to the quantum data compression. The objective of the Quantum data compression researches is to help the engineers to build efficient quantum machines, and leading the computer scientists to invent algorithms to best utilize the quantum machines resources. This paper briefly surveys the recent methods and approaches used to achieve Quantum data compression and the recommendations for further development of Quantum data compression and the focus will be related to the computer science specially that uses the neural network to compress the data.

Keywords: Quantum computing, Quantum concepts, Quantum computing history, Quantum data compression, Quantum neural networks, Quantum Auto-encoder

1. Introduction

Data compression, is the process of reducing the amount of data needed for the storage or transmission of a given piece of information. Compression of classical data is a simple procedure that allows a string of information to take up less space in a computer's memory.

But a Quantum information is carried in the form of “qubits” superposition of quantum states that can be both one and zero in different proportions at the same time. Reading (measuring) a qubit converts it into a classical bit with a definite one or zero value (collapse process), no longer in a combination state. For example, measuring roughly $N/2$ ones and $N/2$ zeros may indicate that we’ve input equal parts of ones and zeros in qubits, which is wrong because the qubit data contain much information than when they are collapsed or measured. But taking a single such qubit and reading it N times would produce the same answer over and over, because after the first reading, the qubit becomes classical. In other words, the N -qubit state contains more information than the classical reading procedure can convey, so the classical compression method will not work. However, a mathematical argument shows that the information contained in a set of N identical qubits can in principle be carried by many fewer qubits; the number is approximately the logarithm of N .

¹ Institute of Statistical Studies and Research, Cairo University

² Faculty of Computers and Information Cairo, University

2. Quantum computing background & history:

A quantum computer is a physical computational device designed to use the principles of quantum mechanics to execute algorithms which are fundamentally impossible for any computer which only uses classical physics, so in order to understand the quantum computing we should learn about some concepts of the quantum mechanics which don't occur in the standard, classical, physics and a brief history about the quantum computing. [See: Yanofsky N.S. et al. (2008), Kaye P. et al. (2010)].

2.1 The Concepts of Quantum Computing:

Quantum mechanics: subfield of physics which studies phenomena at the micro level [See: Yanofsky N.S. et al. (2008), Kaye P. et al. (2010)].

Superposition: If a classical system can be in state of 0 or state 1, a quantum system can be in a "mixture" of the two states. But If we try to measure its state, we see either 0 or 1, probabilistically. The superposition is actually the symbols of both 0 and 1 and all the points in between them [See: Yanofsky N.S. et al. (2008)].

Collapse: a state that an electron or photon takes once we try to measure its state $|0\rangle$ or $|1\rangle$ [See: Yanofsky N.S. et al. (2008), Kaye P. et al. (2010)].

Entanglement: to simplify the definition, we can make two atoms closely related by applying outside force to one atom. When this happens, the second atom takes on the properties of first atom. And from the second atom, applying some quantum physics methods, we can know the value of our atom of interest [See: Yanofsky N.S. et al. (2008), Kaye P. et al. (2010)].

Uncertainty: founded by Heisenberg and It states that we cannot simultaneously assign values of the pairs of the position and momentum, with full precision [See: Yanofsky N.S. et al. (2008), Kaye P. et al. (2010)].

Qubit: the basic building-block of quantum computers, so it's a number of elemental particles such as electrons, ions or photons can be used with either their charge or polarization acting as a representation of 0 and/or 1 so the qubit can exist in superposition, which means that it can exist in multiple states at once [See: Yanofsky N.S. et al. (2008), Kaye P. et al. (2010)].

2.2 History of QC:

In 1980, Benioff pointed out a possibility of constructing a computer based on quantum mechanics [See: *Benioff P., (1980)*]. He proposed a quantum mechanical model of computers within the framework of Turing Machine, which is a standard model of computation. As a consequence, it cannot exceed the computational power of a classical computer. Thus, his model cannot be interpreted as a new model for quantum computers.

In the same year, Yuri Manin proposed the idea of a quantum computer with his book "Computable and Un-computable".

In 1982, Feynman, a Nobel prize winner, Feynman pointed out that some kind of quantum mechanical phenomena cannot be effectively simulated by traditional classical computers and claimed that "computation" for simulating physics can be performed more effectively by using quantum mechanical phenomena. According to Feynman, quantum computers

should have a computational mechanism which obeys the laws of quantum mechanics. Additionally, Feynman introduced some examples of physical phenomena which can be interpreted as a computational model for quantum computers [See: Feynman R. P., (2011)].

In 1985, Deutsch proposed a computational model called the “quantum computer” which reformulates a Turing Machine and has the computational power equivalent to quantum Turing Machine. [See: Deutsch, D., (1985)] it is the first computational model for quantum computing. He also showed a theory of quantum gate, which is another model of quantum computer, in 1988. [See: Deutsch, D., (1989)].

In 1993, Bernstein and Vazirani proposed a universal quantum Turing Machine which generalizes Deutsch’s quantum Turing Machine [See: Bernstein E. et al., (1993)].

In 1994, Shor demonstrated a polynomial-time quantum algorithm for prime factorization and discrete logarithm [See: Shor, P., (1994),(1997)]. He provides a quantum algorithm for prime factorization is called Shor’s algorithm. It is observed the algorithm justifies a strong possibility of quantum computers. This is because prime factorization is known as one of the intractable problems which is the underlying base of RSA code. Shor’s result proves that there is a possibility of breaking the RSA code if a quantum computer is existing and prime factorization is performed vastly by quantum computer. The result showed the importance of quantum computers.

In 1996, Grover proposed quantum search algorithm in Grover [See: Grover, L., (1996)]. Grover showed the quantum algorithm which can search data in unstructured database in the order of speed up an unstructured search problem quadratically, but its uses extend beyond that; it can serve as a general trick or subroutine to obtain quadratic run time improvements for a variety of other algorithms. This is called the amplitude amplification trick; the algorithm is called the Grover’s algorithm.

In 1998, Ömer implemented a programming language for quantum computers called “QCL” at Technical University of Vienna [See: Ömer, B., (2003)]. QCL is a procedural language like C language and can be used for implementation and simulation for various quantum algorithms, so studying the quantum computing in software level became easier.

In 1998, Gershenfeld and Chuang at MIT developed a 2-qubits quantum computer based on NMR (Nuclear Magnetic Resonance) [See: Gershenfeld, N. et al. (1997)]. This is the first actual implementation of a quantum computer. This implementation inspired IBM to implement its first 7-qubits NMR quantum computer and implemented Shor’s Algorithm, thus showing the effectiveness of NMR for the implementation for quantum computers.

3. Quantum Data Compression Methods

In April 2002, Patrick Hayden, Richard Jozsa and Winter [See: Hayden P., (2002)], studied the visible compression of a source of pure quantum signal states or, more formally, the minimal resources per signal required to represent arbitrarily long strings of signals with arbitrarily high fidelity. Their result was a formula to express the trade-off curve $Q^*(R)$ between quantum and classical resources in term of single letter optimization [See: Hayden P., (2002)].

In 2003, Vaccaro [See: Vaccaro J. A., (2003)] in his work touched some basics concepts about the Quantum theory and the impersonation of the Qubit so they described the differences between classical information theory comparable to quantum information theory, so Vaccaro described the representation of a string of letters in the classical information

theory, each letter can define the property of classical message which is any letter can be represented only as one of the binary states **0** or **1**. On the other hand a Quantum message is not so restricted to that but a quantum memory element, (the qubit), can be in a state representing a superposition of **zero** and **one**.

Since the photon is a particle of light so it needs a device called photodetector that can detect individual photons. Photodetectors can be operated to give a voltage spike once a photon is detected [See: Vaccaro J. A., (2003)]. Vaccaro, described some devices for manipulating photons that consists of mirrors and beam splitters and combiners. A typical mirror consists of a reflective coating on a glass slide. In the absence of a coating the light is simply transmitted through the glass. In the intermediate system, a coating can be made sufficiently thin that, for example, 50% of a dense beam of light is reflected and 50% is transmitted. Such a semi-silvered mirror is called a 50-50 beam splitter, since it splits the beam equally into two. So if a single photon directed at a 50-50 beam splitter that path the photon takes should be measured, by placing a photodetector in each of the reflected and the transmitted paths as shown in Figure 1-a. The probability of founding the photon in either transmitted or reflected path will be equal. Vaccaro labels the reflected path as a logical **0** and the transmitted path as a logical **1**. This simple demonstration illustrates how a photon can represent a bit of information [See: Vaccaro J. A., (2003)].

On the contrary side, merging the reflected and transmitted beams using mirrors and a beam combiner, as shown in Figure 1-b. If the lengths of the two possible paths 0 and 1 are such that the crests of both waves arrive simultaneously at the beam combiner, the photons will emerge in the downwards direction as shown in the figure. but, if the crests of one wave coincides with troughs of the other at the beam combiner, the photons will emerge in the upwards direction indicated by the dotted line in the figure. By slightly changing the relative path lengths, the photons can be switched from the downwards to the upwards path. So, photons once arrived at the beam combiner, they already have the information of both path lengths which indicates that a single photon exists simultaneously in both paths. So while the photon is between the beam splitter and beam combiner, the photon represents both logical values 0 and 1, Superposition state [See: Vaccaro J. A., (2003)].

Vaccaro, introduced an implementation to compress two qubits, and three qubits scheme the methods are presented in the following paragraphs.

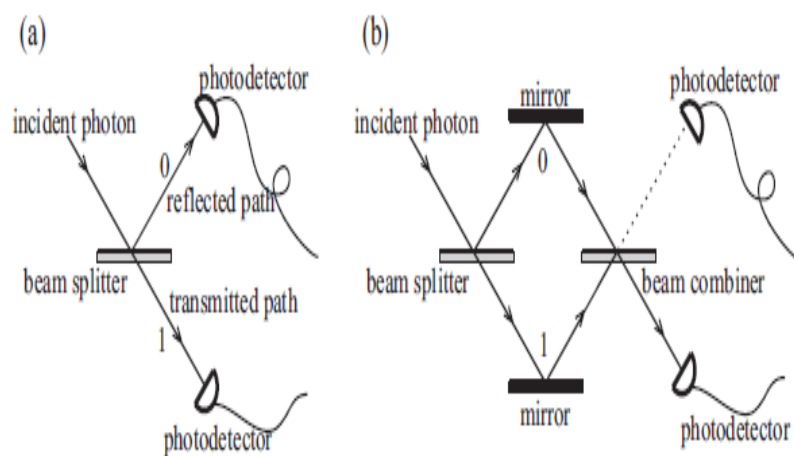


Figure 1: A single photon representing 1 qubit [See: Vaccaro J. A., (2003)]

Compressing 2 Qubit Scheme: It requires the photon to be in a superposition of 4 paths, which had been achieved by splitting the **0** and **1** paths into two further paths for each as shown in Figure 2-b. The photodetector, here, is used to compress a **4-states** system into a **3-state** system to measure the presence or absence of a photon in path “**11**”. The detection of a photon indicates the latter case and an extra photon is switched into path **00** to represent the coding of the state, as indicated by the dotted line in Figure. 2b. The decoding is produced simply by making available an empty path representing the state [See: Vaccaro J. A., (2003)].

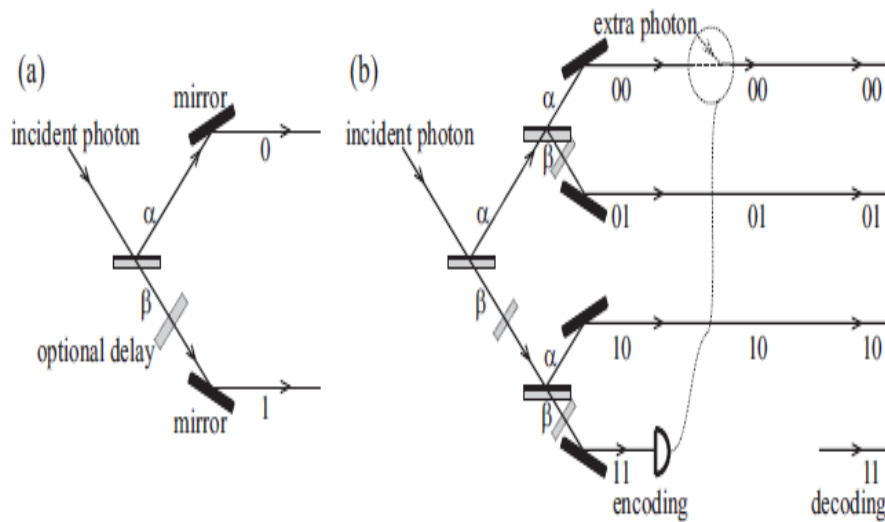


Figure 2: Implementing (a) a single letter state and (b) the 2 qubits scheme [See: Vaccaro J. A., (2003)].

Compressing 3 Qubit Scheme: It can be achieved by further splitting each of the 4 paths into two with the same idea, as before. A single photon in this system is in a superposition of 8 paths. The unitary transformation is produced, simply, by interchanging the two paths **100** and **011** as shown in Figure 3. If one of the photodetectors records a photon, an extra photon is switched into the optical path **00**.

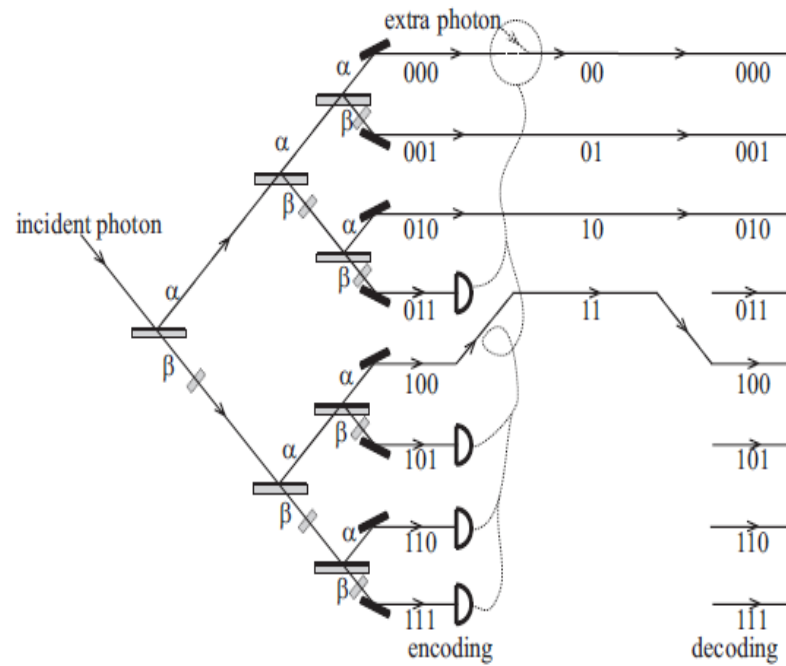


Figure 3: Implementing the 3-qubit scheme [See: Vaccaro J. A., (2003)].

Vaccaro, also performed an experimental implementation of the three-qubit scheme, in which he makes use of an extra degree of freedom available to photons was ignored up till now. The electromagnetic waves comprising the photons oscillate transversely to the direction of travel. The orientation of the oscillations is called the polarization of the photon and gives rise to an intrinsic polarization qubit. Thus a 3-qubit

system can be constructed from two path qubits (i.e. 4 paths) and 1 polarization qubit (2 orthogonal polarization directions). The extension of the 2-qubit and 3-qubit optical schemes to blocks of larger K is straightforward requiring $2K$ optical paths. However, the scheme quickly becomes experimentally difficult to perform owing to the degree of control required over the many optical paths.

Rather than using a single photon to represent a multiple qubit block, a more scalable system comprises one photon for each qubit. However, the various operations needed for coding and decoding require nonlinear interactions between the photons. Such nonlinear interactions between individual photons have not yet been produced experimentally, although there are proposals for engineering them using available technology [See: Vaccaro J. A., (2003)].

In 2003, Miguens compared between the classical information theory and the Quantum information theory. Quantum Information Theory is fundamentally richer than Classical Information Theory, one of the central differences is the nature of the information itself.

She agreed with Jozsa [See: Jozsa R., (2003)], who introduced an overview of the main properties of quantum information. In his opinion, the classical compression scheme cannot be used to compress quantum data because of quantum systems behavior [See: Joana I.L. Miguens (2003)].

Miguens, therefore, started to formulate a quantum encoding scheme, and introduced three methods to improve the encoding scheme for quantum data compression, which are:

1. Brute Force (BF), presents only significant advantages for small amounts of data, but has otherwise a very high complexity time.
2. An improved BF scheme, where the complexity time is decreased by taking benefit of the advantages of quantum algorithms.
3. An adapted Algorithm, based on a different way of ordering probabilities which offers the best complexity time. This algorithm contrary, to the previous two methods, does not guarantee conversion to an optimal solution.

In 2007, Reif and Chakraborty [See: Reif J. H. et al. (2007)], presented an algorithm for optimal quantum compression/decompression, using $O(n(\log^4 n) \log \log n)$ elementary quantum operations. They figured out that the quantum methods cannot increase the bandwidth for transmission of classical information. However, entangled quantum states can be compressed much more than it is possible via classical lossless compression. Following Schumacher, they assumed the existence of a finite quantum state ensemble which is a mixed state consisting of a finite number of qubit states and their associate results, providing the first quasi-linear time algorithm for asymptotically optimal (in size and fidelity) quantum compression and decompression. Reif and Chakraborty also, outlined the quantum gate array model to bring about this compression in a quantum computer.

Also, in 2007 an Adaptive Quantum Lossless Compression algorithm was presented by Al-Daoud [See: Al-Daoud E., (2007)]. The main idea of the proposed algorithm is that the corresponding probabilities of the symbols are, initially assumed to be identical, then, after each iteration; all the corresponding probabilities are updated. These probabilities will converge to the actual distribution after a few iterations. Moreover, the complexity of the proposed algorithm can be reduced, if the general properties of the given data are considered.

In 2010, physicists Plesch and Bužek [See: Plesch M. et al. (2010)], realized that, while it is not possible to compress quantum data to the same extent as classical data, some compression can be achieved. As long as the quantum nature of a string of identically prepared qubits is preserved, it should be possible to feed them through a circuit that records only their probabilistic natures. Such a recording would require exponentially fewer qubits, and would allow a user to easily store the quantum information in a quantum memory, which is currently a limited resource. Then at some later time, the user could decide what type of measurement to perform on the data.

In Oct 2014, Rozema [See: Rozema L. A., (2014)], have devised a way to put qubit compression into practice. In the simplest case, the compression method puts the information contained in three identical qubits into two qubits that are not identical. Information common to all three qubits is encoded only once in the pair instead of redundantly in all three. Physically, compressing the qubits means making them interact to generate two new qubits with the properties prescribed by the mathematical recipe. Figuring out how to do this was the hard part of the problem. As in any manipulation of quantum information, the challenge was to perform actions on the qubits without destroying their quantum nature.

In 2016, Wan [See: Wan K. H., (2016)], proposed a Quantum generalization of Classical feedforward neural network. Since the classical neural networks are involved in many applications such as decision making, Image recognition, Data science, data compression and AI, Wan, thought about generalized a Quantum model to mimic the classical neural network. He took in his consideration the challenge of the classical-

quantum clash. The clash between the nonlinear, dissipative dynamics of neural network computing and the linear, reversible dynamics of quantum computing, [See: Schuld M.,(2014)].

Wan believed that there are promising results concerning the quantum versions of recurrent neural networks, wherein neurons talk to each other in all directions rather than feeding signals forward to the next layer. He had the purpose of implementing quantum simulated annealing [See: Wiebe N.,et al. (2014), Heim B. et al (2015), Lechner W. et al. (2015), Garnerone S. et al. (2012)], using the superposition property, and he aimed to find a way to compress data encoded in quantum superposition to minimal number of qubits [See: Wan K. H., (2016)].

Wan overcomes the mentioned challenge, namely, the Classical-quantum clash, by utilizing feedforward perceptron neural network, auto-encoder, that converts every single neuron to be reversible, then generalized those reversible neurons to be quantum reversible (unitary) [See: Wan K. H., (2016)]. To evaluate how close, the outputs are to the desired outputs, he used a global gradient descent cost function to his generalization [See: Wan K. H., (2016)]. He, described the classical neural network with two inputs, and one output, then he assigned weights to the inputs, so if the summation of the weighted inputs is above a predefined threshold the output is **1**, and **0** otherwise.

In Figure 4, Wan described a classical neuron that takes two inputs in_1 and in_2 with input weights $w_{j1}^{(i)}$, $w_{j2}^{(i)}$, respectively. The output of j^{th} neuron in the i^{th} layer of the network is $a_j^{(i)}$, and it fires the function $z_j^{(i)}$ of the weighted input, where

$$z_j^{(i)} = \sum_{k=1}^n w_{jk}^{(i)} * a_k^{i-1} \quad (1)$$

This function relates the output at layer i with the weighted output of layer $i-1$, it is called the *activation function*, which has, most commonly, been a Heaviside step function or a sigmoid [See: Wan K. H., (2016)]. The Heaviside activation function gives an output of the form:

$$a_j^i = \begin{cases} 1, & z_j^{(i)} > 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Using the above description, Wan generalized this classical neural network to a quantum neural network by converting this irreversible neuron to a reversible gate and finally to unitary gate [See: Wan K. H., (2016)].

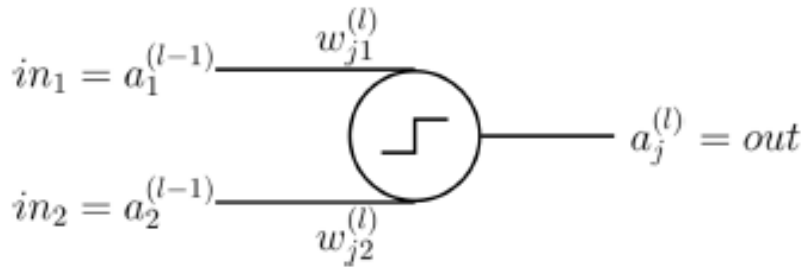


Figure 4: A classical neuron taking two inputs [See: Wan K. H., (2016)]

The following is the procedure for generalizing the neural network to quantum neural network.

Step 1: converting from Irreversible to reversible:

For an n-input classical neuron having $(n_1, n_2, \dots, n_n) \rightarrow \text{out}$, create a classical reversible gate taking $(n_1, n_2, \dots, n_n, 0) \rightarrow (n_1, n_2, \dots, n_n, \text{out})$

Such an operation can always be represented by a permutation matrix [See: Muthukrishnan A., (1999)]. This is a clean way of rendering the classical neuron reversible. The extra dummy input bit is used to make it reversible [See: Feynman R. P., (1986)]; in particular, the implementation of a reversible neuron is by adding an extra bit to the 2-bits model.

Step 2: Reversible $\rightarrow$ unitary:

Generalize the classical reversible gate to a quantum unitary.

Thereafter, Wan described a compression device, using auto-encoder neural network, and showed how the input is propagated between the layers to compress two inputs into one hidden layer, the bottleneck layer.

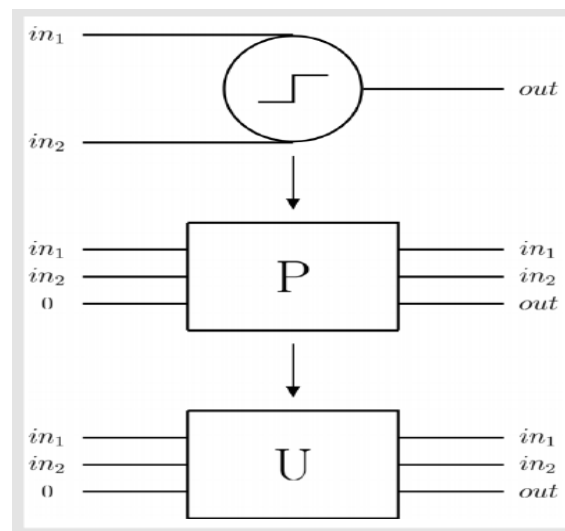


Figure 5: Generalization of classical irreversible to quantum reversible computation [See: Wan K. H., (2016)].

He, also, illustrated how to use the output layer to train the network, the compression device is shown in Figure 6, inside the blue dotted frame.

A cost function is used to alter the weights after each propagation to make the network's output closer to the desired output. He applied step two to a quantum case by generalizing the neurons labelled 1, 2 and 3 in Figure 6, into unitary matrices U_1 , U_2 and U_3 , with the addition of a 'fan-out' gate. The 'fan-out' gate is, actually, distributing information about the input state into several output qubits.

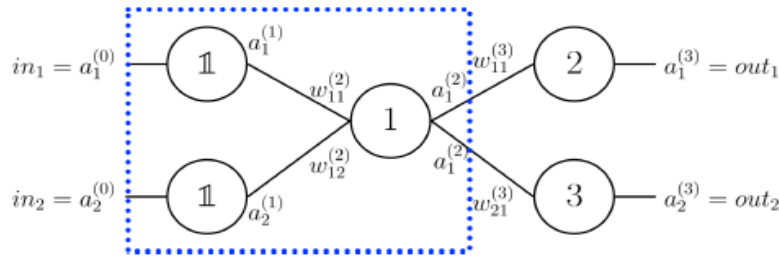


Figure 6: A classical auto-encoder taking two inputs [See: Wan K. H., (2016)].

Figure 7, illustrates a Neural network implementing a quantum auto-encoder that can accommodate two entangled input qubits. The blue box represents the quantum compression device after training.

The training rule is, therefore, the same but replacing the weight $w_{jk}^{(i)}$ with a general parameter $\alpha_{j_1, \dots, j_N}$ of the unitary U_N [See: Wan K. H., (2016)].

Wan's work inspired Romero [See: Romero J., (2017)] to apply a simplified model of auto-encoder using the swipe gate in order to find the best unitary $U^{\bar{p}}$ which preserve the input state of the circuit. Figure 8, illustrates Romero's proposal, where he employed a swap gate to trace over the B' system instead of tracing over the B system.

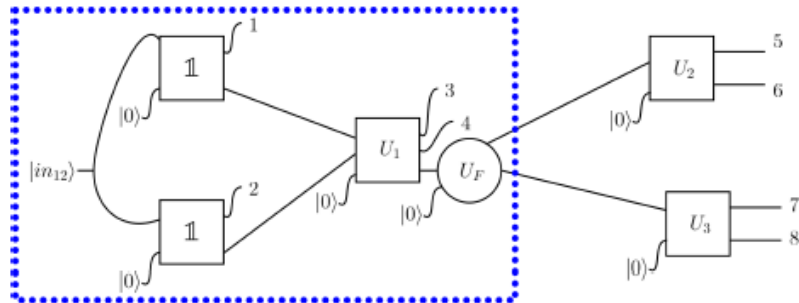


Figure 7: Quantum Auto-encoder [See: Wan K. H., (2016)]

In his work Romero recognized patterns that are beyond the capabilities of classical machine learning. His motivation for a quantum auto-encoder is that; such a model allows to perform machine learning tasks for quantum systems without the exponentially costly classical memory. Which is very beneficial in quantum data applications including, dimension reduction of quantum data, and the quantum simulators.

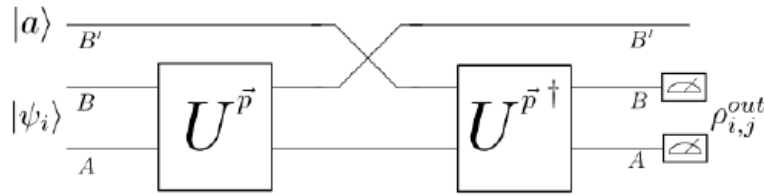


Figure 8: The proposed Quantum auto-encoder circuit by Romero [See: Romero J., (2017)].

In Feb 2017, Yanga [See: Yanga, (2017)], studied the compression of arbitrary parametric families of n identically prepared finite-dimensional quantum states, in a setting that can be regarded as a quantum analogue of population coding.

In June 2018 Boström and Felbinger [See: Boström K., (2002)], developed a general framework for variable-length quantum messages in close analogy to the classical case and show that lossless compression is only possible if the message to be compressed is known to the sender. They showed that it is possible to reduce the number of qubits passing through a quantum channel even below the von-Neumann entropy by adding a classical side-channel.

4. Conclusions

This paper, introduced the concept of quantum computing and summarized the methods for manipulating qubits. Most of the mentioned works depend on ideas from physics or mathematics to show the possibility of compressing the quantum data. One of the methods, that uses the physics to engineer a compression device, was presented in details, just to imagine what and how the photons interact with the data and how could it to be measured, on the other side as computer science approach. Wan [See: Wan K. H., (2016)] build his work using deep learning based on neural network techniques. His work is promising as it provides an auto-encoder neural network that encodes the quantum data for dimensional reduction. The network should be trained to recreate the signal at encoder layer, which necessitates reversibly compressing it to the number of neurons in the bottle-neck layer. The post-bottleneck part of the network, decoder, can be discarded after training the network and the compressed output taken from the bottleneck directly as the desired compressed data.

An important note is that, the Heaviside function, is only useful within single-layer network and only handles linearly separable data. The desired quantum compression level, however, requires multilayer network, even, using it as a step function is non-differentiable at $x = 0$ and its derivative is 0 elsewhere. This means gradient descent won't be able to make progress in updating the weights and backpropagation will fail.

To overcome these two problems, the authors suggest to enhance this model by replacing the Heaviside function with sigmoid function as it doesn't have these limitations. It is useful in the training phase as the backpropagation in a multilayers network can be considered to achieve the maximum level of lossless compression of quantum data.

References

- Yanofsky N.S., Mannucci M.A. (2008), Quantum Computing for Computer Scientist, Cambridge university press, 978-0-521-87996-5
- Kaye P., Laflamme R., Mosca M., (2010), An Introduction to Quantum Computing, OXFORD University Press, 978-0-19-857049-3
- Benioff P., (1980), The computer as a physical system. Journal of Statistical Physics 22, 563–591
- Feynman R. P., (2011), Feynman lectures on Physics, Vol III, Basic Books 978-0465025015
- Deutsch, D., (1985), Quantum theory, the Church-Turing principle and the universal quantum computer. Proc. of the Royal Society A 400, 97–117
- Deutsch, D., (1989), Quantum computational networks. Proc. of the Royal Society A 425, 73–90
- Bernstein E., Vazirani U., (1993), Quantum complexity theory. In: Proc. of the 25th Annual Symposium on the Theory of Computing, pp. 11–20
- Shor, P., (1994), Algorithms for quantum computation: Discrete logarithms and factoring. In: Proc. of the 35th Annual Symposium on Foundations of Computer Science, pp. 124–134
- Shor, P., (1997), Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM Journal of Computing 26, 1484–1509, Expanded version of [8]
- Grover, L., (1996), A fast quantum mechanical algorithm for database search. In: Proc. of the 28th ACM Symposium on Theory of Computing, pp. 212–219
- Ömer, B., (2003), Structured Quantum Programming, Institute for Theoretical Physics, Vienna University of Technology. Ph.D. dissertation
- Gershenfeld, N., Chuang, I., (1997), Bulk spin resonance quantum computing. Science 275, 350– 356
- Hayden P., (2002), Trading quantum for classical resources in quantum data compression, Journal of Mathematical Physics Vol.43,
- Vaccaro J. A., (2003) Quantum Data Compression, A. Albrecht and K. Steinhofel (Eds.): SAGA 2003, LNCS 2827, pp. 98–107, 2003. Springer-Verlag Berlin Heidelberg
- Jozsa R., (2003),” Illustrating the concept of quantum information”, IBM J. Res.Dev. Article for Charles H. Bennett 60th Birthday Symposium. Electronically available at quant-ph/0305114.
- Joana I.L. Miguens, Quantum Data Compression, November 9, 2003

Reif J. H. and, Sukhendu C., (2007) Efficient and exact quantum compression, Science direct, Information and Computation 205 (2007) 967–981, Received 2 June 2005; revised 11 July 2006, Available online 2 February 2007

Al-Daoud E., (2007) Adaptive Quantum Lossless Compression, Journal of Applied Science 7(22): 3567-3571, ISSN 1812-5654

Plesch M., Bužek V., (2010), Efficient compression of quantum information, arXiv:0907.1764v2 [quant-ph] 2 Feb 2010

Rozema L. A., (2014), Quantum Data Compression of a Qubit Ensemble, 0031-9007=14=113(16)=160504(5), American Physical Society

Schuld M., Sinayskiy I., and Petruccione F., (2014), The quest for a quantum neural network, Quantum Information Processing 13, 25672586

Wiebe N., Kapoor A., and Svore K. M., (2014), Quantum deep learning, unpublished.

Heim B., Rnnow T. F., Isakov S. V., and Troyer M., (2015), Quantum versus classical annealing of Ising spin glasses, Science 348, 215{217}

Lechner W., Hauke P., and Zoller P., (2015), A quantum annealing architecture with all-to-all connectivity from local interactions, Science Advances 1

Garnerone S., Zanardi P., and Lidar D. A., (2012), Adiabatic quantum algorithm for search engine ranking," Phys. Rev. Lett. 108, 230506

Wan K. H., (2016), arXiv:1612.01045v1 [quant-ph]

Muthukrishnan A., (1999), Classical and quantum logic gates: An introduction to quantum computing, seminar notes, unpublished

Feynman R. P., Quantum mechanical computers Found, (1986), Physics 16, 507531

Romero J., (2017), Quantum auto-encoders for efficient compression of quantum data, arXiv:1612.02806v2 [quant-ph]

Yanga, (2017), Compression for quantum population coding, arXiv:1701.03372v3 [quant-ph].

Boström K., (2002), Lossless quantum data compression and variable-length coding, arXiv:quant-ph/0105026v2

The Leap from Digital Quantum Simulators to Quantum Supremacy

Hisham Nabil¹, Fatma A. EL-Licy¹, Akram Salah²

ABSTRACT

During the past five years, an impressive progress has been made toward achieving the quantum supremacy, the business of computer hardware and software industry was funding the quantum computing research to develop the quantum hardware using several approaches. This paper focus on the current state of quantum computing and discusses in more details the recent progress in the quantum computing simulation, and how it is affected by the large organizations' work during the past five years. The paper will discuss a brief background about the quantum computing principles, also the difference between the universal quantum computer and the quantum computer based on the simulated annealing.

Keywords: Quantum computing, Quantum simulators, Quantum computing history, Quantum computing industry, Quantum programming languages.

1. INTRODUCTION

The field of quantum computing has been developed quickly during the last five years. The quantum computing researches are running in different ways simultaneously toward the quantum supremacy. The key players in the business of computer hardware and software industry are, now, allocating budgets to support the research and development on the quantum computing. The quantum computing as a theory was formulated by several researches during the early eighties of the last century. Paul Benioff, Yuri Manin, Richard Feynman, and David Deutsch [See Feynman R.P. (1982), Benioff P. (1980), Manin, Yu. I. (1980), Deutsch D. (1985)], were fascinated by the quantum theory as a scientific idea. It describes how to build a computing model based on the quantum mechanics principles, like the superposition, spinning, measurement, and entanglement.

Many digital quantum simulators have appeared after that, and many surveys have been done on the progress in quantum simulators and quantum programming languages. In parallel lots of experiments are performed to build a real quantum machine.

This paper will walk through the main principles of quantum physics that led to the idea of quantum computing. The main components of the quantum computing programming model, will be presented as well as the current impressive advances in the quantum computing hardware, simulators, and programming languages. The light will, also, be spotting on some important related definitions.

¹ Institute of Statistical Studies and Research, Cairo University

² Faculty of Computers and Information Cairo, University

2. QUANTUM COMPUTING BACKGROUND

Most of the hot topics related to computers are discussing the modern capabilities that computers are bring to human life. Yet, it is rarely happened to discuss the things that today's computers are not capable of doing, or, to address the limitations of the current computational model that exploit the physical properties of the semiconductors and transistors that act as the building blocks for the classical computers architecture.

Quantum computing is the discipline of science that went beyond the classical physics limitations and expand the computing capabilities by exploit the fundamentals and principles of quantum mechanics [See Yanofsky N.S. et al. (2013), Kaye P. et al. (2010)].

Limitations of the Digital Machines

The digital machines are limited, when it comes to, solve problems that needs simulation of data with a size of memory that, exponentially, increases by the increase of the input length. Examples of problem that shows such limitations, include:

- The factoring problem for a number into primes $M = p * q$, when both numbers (p, q) are very large prime numbers,
- The calculation of CaF Calcium monofluoride bond length, the actual length is 1.967, while, the calculated one by the digital quantum simulator is 4.079,
- The calculation of the Na₂ Sodium diatomic bond length, the actual length is 3.079, while, the calculated one by the digital quantum simulator is 2.379.

Quantum computing has many real life possible applications once we get into real quantum computing universal machine. Its applications include medicine research industry, cryptography, telecommunication, and solving optimization problems [See Yanofsky N.S. et al. (2013), Kaye P. et al. (2010)].

Quantum Mechanics

Quantum Mechanics is a fundamental branch of physics, discovered just after 1920, which explains the strange behavior of the waves that behave like particles (especially the higher frequencies waves). It also, describes the particles that behave like waves as in the double-slit experiment by Thomas Young (1773 – 1829).

An important difference between the quantum mechanics and the classical theories of physics is that we cannot use our human sense while understanding the properties of the quantum mechanics. Properties of Quantum mechanics is not like those of classical mechanics represented by the classical image of our universe as a three-dimensional space of geometry, as described by Euclid.

Newton's laws of motion can remind us with the human sense flavor in Physics:

- 1- An object at rest tends to stay at rest and that an object in uniform motion tends to stay in uniform motion unless acted upon by an external force.
- 2- An applied force, F, on an object equals the rate of change of its momentum, p, with time.
- 3- For every action there is an equal and opposite reaction.

In quantum mechanics, though, things are quite different [See Feynman R. P. (1963)]:

- 1- **Photons, the quantization of light:** Einstein showed that the quantization of light energy, explains several other phenomena, like the photoelectric effect,
- 2- **The Uncertainty principle:** One cannot know both where something is and how fast it is moving. The uncertainty of the momentum and the uncertainty of the position are complementary, and the product of the two is bounded by a small constant (Planck's constant),
- 3- **Wave-particle duality:** Every elementary particle exhibits the properties of not only particles, but also, waves,
- 4- **Quantum entanglement:** Is a physical phenomenon that occurs when pairs- or groups of- particles are generated or interact in ways such that the quantum state of each particle cannot be described independently, instead, a quantum state may be given for the system. Measurements of physical properties such as position, momentum, spin, polarization, ...etc., performed on entangled particles are found to be appropriately correlated.

The Interference and Superposition

The double-slit experiment by Young was the initial stone that breaks the classical physics world of principles. One shines light at a boundary with two slits that are very close to each other. The pattern of the light to the right of the boundary will have certain regions that are dark and certain other regions that are bright, as depicted in Figure 1.

The reason why there are regions on the screen with no light is that light waves are interfering with each other. Light is propagating as a single wave from its source.

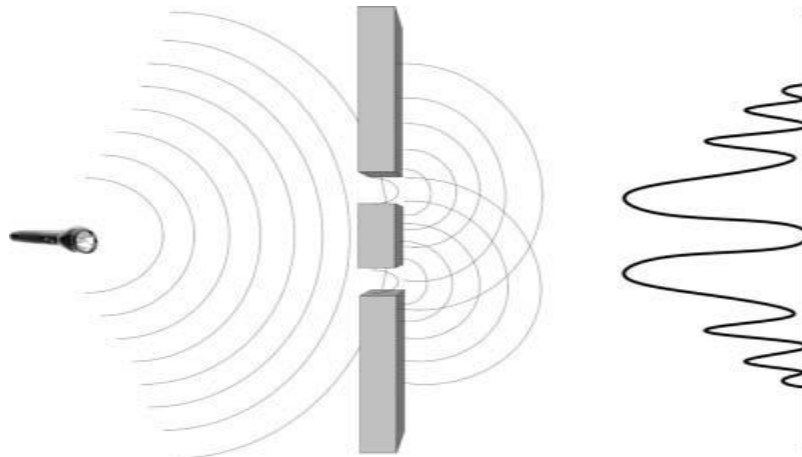


Figure 1: Young's double-slit experiment [See Yanofsky N.S. et al. (2013)]

The two slits cause this wave to split into two independent ones, which can then interfere with each other when reaching the screen. Some regions are going to be darker, others are going to be brighter, depending on whether the two waves are in phase (positive interference) or out of phase (negative interference).

When one of the slits were closed off, there was no splitting and therefore no interference pattern, whatsoever, occurs, as shown in Figure 2.

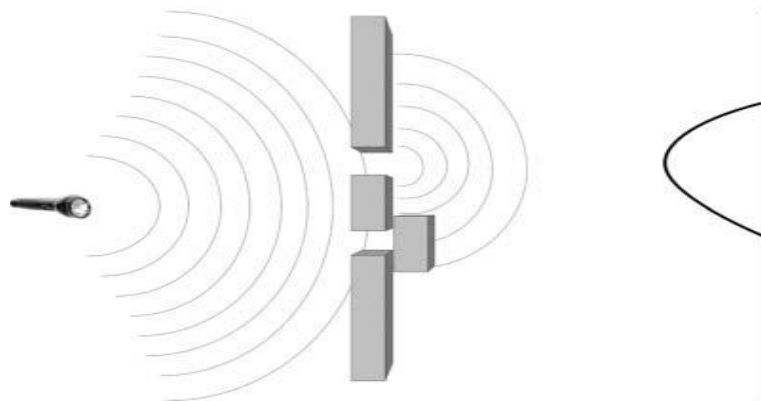


Figure 2: Young's double-slit experiment with one slit closed [See Yanofsky N.S. et al. (2013)]

The double-slit experiment can be done with just one photon at a time. Rather than spotting patterns of lighter or darker light on the screen, we are now looking for which region is likely for the single photon to land. The same pattern can then be viewed as describing the probability for a certain region to get hit by the photon. The fact that the photon is producing the interference pattern even if it is one photon, or electron, means that, photon is able to be passing through both slits at the same time, which is called the *superposition* of states. The natural question then is, if there is a single photon why would there be any interference pattern? Yet, experiments have shown that such a pattern is there. The photon is a true chameleon: sometimes it behaves as a particle and sometimes as a wave, depending on how it is observed.

The double-slit experiment is not only about light: one can perform it equally well with electrons, protons, and even atomic nuclei, and they will all exhibit the same interference behavior. Once again, this clearly indicates that the rigid distinction between waves and particles as a paradigm of description of the physical world is untenable at the quantum level [See Yanofsky N.S. et al. (2013)].

The Theory of Quantum Computing

The theory of quantum computing was built on top of some important concepts:

- 1- **The Qubits and the Quantum States:** the main idea starts with representing data using the quantum bit (Qubit) instead of the classical normal Zero or One-bit representation. The classical representation let the zero or one bit consume one bit in the classical digital memory. A quantum bit or a qubit is, however, a unit of information describing a two-dimensional quantum system. It can be represented by a vector of two complex numbers [See Yanofsky N.S. et al. (2013)].
- 2- **Observables:** because there is no straightforward observable, like the mass in classical physics, some imitation, for each physical observable, were used which is corresponding to a Hermitian operator.

- 3- **Measuring:** The measuring here is to get the observable value, which make it collapse to one final state and lose the superposition.
- 4- **Quantum Gates:** A quantum gate is simply an operator that acts on qubits. Examples on quantum gates are (Phase shift, unitary gate, Hadamard gate, ...etc.)
- 5- **Quantum Algorithms:** Algorithms are often developed long before the machines they are supposed to run on, this was the case in quantum algorithms like (shore's factoring algorithm, Grover's search algorithm, ...etc.)

3. QUANTUM SIMULATORS AND PROGRAMMING LANGUAGES

Complex numbers can represent the possibilities of the particle to be in a superposition state instead of a single state, they, also, may cancel each other out when added. This state has a well-defined physical meaning in quantum mechanics (and in classical wave mechanics as well). This is what is called interference and it is one of the most important concepts in quantum theory.

The main issues regarding the simulator is the exponential growth of the resources needed to represent quantum data [See Yanofsky N.S. et al. (2013)]. The use of matrices and vectors of complex numbers to represents the quantum states is essential to differentiate between the normal probabilistic programming and quantum programming. In the quantum programming, the states may be canceling each other's, so for example if we're using the real probabilities to represents the quantum computing states we'll find that $(p_1 + p_2) \geq p_1$ and $(p_1 + p_2) \geq p_2$, this is for any p_1 and p_2 belongs to R , but if we're using the complex numbers instead, this supports the interfering in an elegant way, because if we assume c_1 and c_2 complex numbers, so $|c_1 + c_2|^2$ need not be bigger than $|c_1|^2$ or $|c_2|^2$.

A quantum computer is a physical computational device designed to use the principles of quantum mechanics to execute algorithms which are fundamentally impossible for any computer which only uses classical physics. To understand the quantum computing we should learn the concepts of the quantum mechanics, and a brief history about the quantum computing [See Yanofsky N.S. et al. (2013), Kaye P. et al. (2010)].

The Universal Quantum Computer

A universal quantum simulator is a quantum computer proposed by Yuri Manin in 1980, and Richard Feynman in 1982. Feynman showed that a classical Turing machine would experience an exponential slowdown when simulating quantum phenomena, while his hypothetical universal quantum simulator would not. David Deutsch in 1985, took the ideas further and described a universal quantum computer. In 1996, Seth Lloyd showed that a standard quantum computer can be programmed to simulate any local quantum system efficiently [See Feynman R.P. (1982), Manin, Yu. I. (1980), Lloyd S. (1996)]. This is unlike the quantum annealing (see section 5), which is used to fix one specific type of problems.

Quantum Supremacy

The quantum supremacy will be achieved when the universal quantum computer has been invented and constructed for a fairly large number of qubits. Therefore, it is not achievable to simulate quantum supremacy with the digital simulators. The quantum machine should be universal, so it will be able to run shore algorithm, and any other algorithms. In October 2017, IBM demonstrated the simulation of 56 qubits on a conventional supercomputer, increasing the number of qubits needed for quantum supremacy [See Kim M. (2017)].

The Decoherence Problem

The decoherence is the loss of purity of the state of a quantum system as the result of entanglement with the environment [See Yanofsky N.S. et al. (2013)]. When we have two wave functions that describes two different electrons for example, and let's assume we've succeeded to start these two waves in the same waveform, so they're sharing the same frequency and the same phase difference, we call this the coherence. It is a kind of correlation between two particles in their quantum states, this supports other quantum properties like the interference to take place.

It is important to keep the coherence between particles as long time as possible in order to run a quantum algorithm in a reliable way and for longer time. Therefore, the problem is the decoherence, which is losing the coherence between these particles due to the dynamic and fast changing quantum states. Accordingly, it is essential to have a strong and reliable fault tolerance mechanism to fix the computational errors that might result from the particles' decoherence. It is also one of the reasons behind keeping the IBM quantum processor under 15 millikelvin degree, to slow down the dynamic changes to the maximum, so it gets more reliable, and solve longer algorithms.

4. THE MAJOR STEPS

The D-Wave Quantum Hardware

D-Wave is the first organization that develop a real quantum machine for both research and commercial purposes, it is not a universal quantum machine, yet, but it is based on quantum annealing.

In 2011, D-Wave systems manufactures commercially available quantum annealing (see section 5) based hardware on which it is possible to run algorithms. In May 2013, it was announced that a consortium of Google, NASA Ames and the non-profit Universities Space Research Association purchased an adiabatic quantum computer from D-Wave Systems with 512 qubits [See Jackson J. (2014), Russell J. (2017), Vincent J. (2016)].

In December 2015, Google announced that the D-Wave 2X outperforms both simulated annealing and Quantum Monte Carlo by up to a factor of 100,000,000 on a set of hard optimization problems [See Neven H. (2015)].

D-Wave's architecture differs from traditional quantum computers. It is not known to be polynomial equivalent to a universal quantum computer, therefore, it could not execute Shor's algorithm, because Shor's algorithm is not a hill-climbing process. Shor's algorithm requires a universal quantum computer. D-wave claims only to do quantum annealing [See Marchenkova A. (2016), Heath N. (2018)].

IBM Q and Superconductors

In May 2016, IBM has published its universal quantum computer to be cloud exposed. The Quantum Information Science Kit (Qiskit), Terra provides the foundations for Qiskit. It allows the user to write quantum circuits, easily, while taking care of the constraints of real hardware, a good source of information about Qiskit is the website (<https://qiskit.org>). The project is running by the Science and Solutions department at IBM that has more than 1500 researcher.

Figure 3, shows the **IBM Q** cloud interface where you can simulate the input qubits by the available 5 qubits on the 5 horizontal lines, then apply the gates from the available gates list, the machine run the algorithm several times to apply error handling, and to get more reliable result by calculating the average for the output, it supports several gates for example: (H: Hadamard, X, Y, and Z: bit-flipping and phase-flipping over X, Y, and Z axis, id: identity gate, also there is the measurement gate).

The IBM Quantum machine is based on superconductors that simulates 5 to 20 qubits per chip, the input data passing to the chip using microwave lines, IBM's quantum systems must be kept cold to overcome the decoherence problem. Any heat in the system can introduce errors, which is why quantum computers are designed to operate at temperatures close to absolute zero, exactly on 15 millikelvin degree, which is colder than the vacuum of space. Figure 4, shows how a quantum computer's dilution refrigerator, made from more than 2,000 components, exploits the mixing properties of two helium isotopes to create such a cold environment for the qubits inside, the IBM Q website is a good source of information about the IBM quantum computing experience (<https://www.research.ibm.com/ibm-q>)

The main idea behind the superconductor quantum computers is to use thin layers of nonconducting material sandwiched between two pieces of superconductors, at very low temperature the superconductors' electrons pair up to form a superfluid flowing with no resistance and traveling through the medium as a single, uniform wave pattern, this wave leaks into the insulating middle. The current flows back and forth through the junction much like a ping-pong ball, in a rhythmic fashion. The qubits are implemented through what is now known as the Josephson junction qubit. In this implementation, the $|0\rangle$ or $|1\rangle$ states are represented by the two lowest-frequency oscillations of the currents flowing back and forth through the junction. The frequency of these oscillations is very high, being of the order of billions of times per second [See Yanofsky N.S. et al. (2013)].

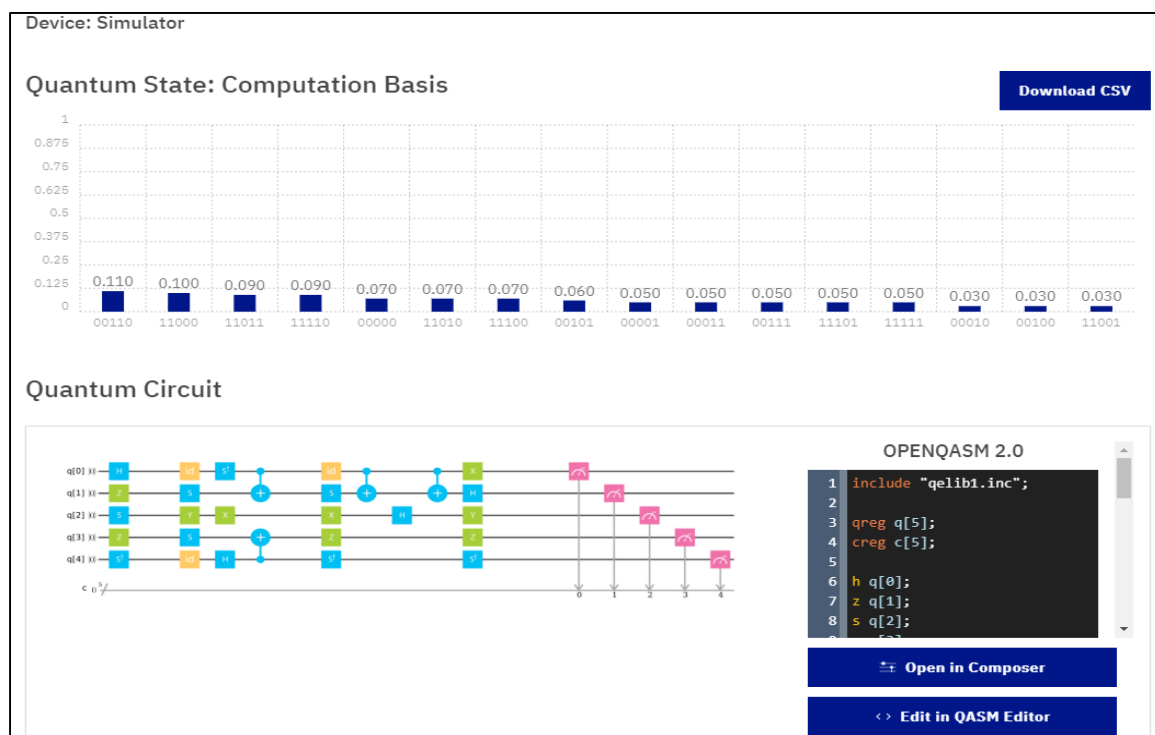


Figure 3: IBM Quantum Machine Cloud Interface

Microsoft Q# and Topological Qubit

In Dec 2017, Microsoft has released the Q# programming language, the language simulates that we write a C# classical program, we call the Q# just when quantum processing needed, so the Q# language provides some functions that represent the quantum gates, for example: (CNOT(), H(): Hadamard, M(): Measurement)

A topological quantum computer is a theoretical quantum computer, the actual implementation is not yet existing, however Microsoft released the simulator to get people familiar with the programming paradigm. The phenomenon of the fractional quantum Hall effect is the one utilized in this new idea of computation, anions are quasiparticles in a two-dimensional space. Anions are neither fermions nor bosons, but like fermions, they cannot occupy the same state. Thus, the world lines of two anions cannot intersect or merge, which allows their paths to form stable braids in space-time. Anions can form from excitations in a cold, two-dimensional electron gas in a very strong magnetic field, and carry fractional units of magnetic flux. In typical laboratory systems, the electron gas occupies a thin semiconducting layer sandwiched between layers of aluminum gallium arsenide.

When anions are braided, the transformation of the quantum state of the system depends only on the topological class of the anions' trajectories (which are classified according to the braid group). Therefore, the quantum information which is stored in the state of the system is impervious to small errors in the trajectories. In 2005, Sankar Das Sarma, Michael Freedman, and Chetan Nayak proposed a quantum Hall device which would realize a topological qubit. In a key development for topological quantum computers, in

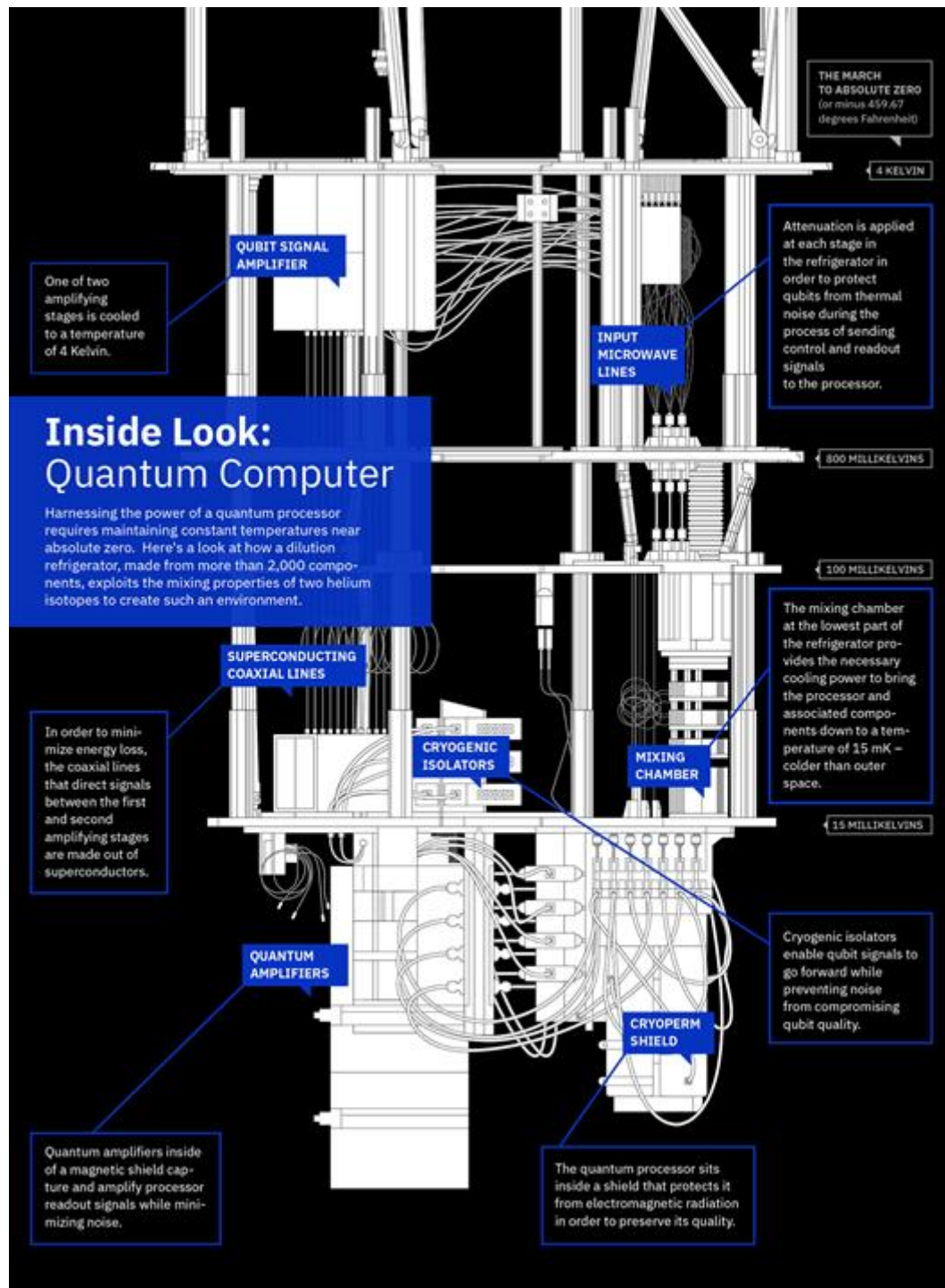


Figure 4: Inside Look - IBM Quantum Machine [See "IBM Quantum Computer" (n.d.)]

2005 Vladimir J. Goldman, Fernando E. Camino, and Wei Zhou were said to have created the first experimental evidence for using fractional quantum Hall effect to create actual anions, although others have suggested their results could be the product of phenomena not involving anions. It should also be noted that non-abelian anions, a species required for topological quantum computers, have yet to be experimentally confirmed [See Willet R.L. (2013), Von Keyserling C. et al. (2015)].

Google Bristlecone

In Mar 2018 Google has released Bristlecone. Google unveiled Bristlecone, a new quantum computing chip with 72 quantum bits, or qubits. John Martinis, who heads Google's effort, says his team still needs to do more testing, but he thinks it's "pretty likely" that this year, perhaps even in just a few months, the new chip can achieve "quantum supremacy." That's the point at which a quantum computer can do calculations beyond the reach of today's fastest supercomputers. The new chip still under testing, and no programming interface is available yet.

5. QUANTUM ANNEALING

To understand the quantum annealing, a different class of search algorithms should be mentioned, which is the local search techniques, resample artificial intelligence tools for solving the optimization problems. One of the most important artificial intelligence topics is the information search and exploration. the systematic search algorithms are designed to explore the search space systematically, by keeping the alternative paths in memory and explore every point. The aim of that is to find the goal along with the path to this goal, this is more suitable when the search space is discrete. Examples on these methods are heuristic search, memory bounded heuristic search, greedy best-first search, A* search, and the use of heuristic functions [See Russell S. et al. (2010)].

Local Search Techniques

The local search techniques are invented to handle the continuous search space, when the purpose is to find the goal regardless of the path to this goal. Also, it can handle the optimization problem that requires to find the local or global minimum or maximum, on the objective function graph. These kinds of problems are very important and usually, requires expensive computing resources. Example of these kind of problems are assignment problem, constraint satisfaction problem, cutting stock problem, traveling salesman problem, scheduling problem, vehicle routing and rescheduling, minimum spanning tree, closure problem, integer programming, and knapsack problem.

This class of search algorithms doesn't worry at all about the paths to the solutions, it operates using a single current state instead of multiple paths. This class uses minimal amount of memory, it can, usually, find a reasonable solution in a continuous state space.

The Hill-Climbing Search

A good example on local search techniques is the hill-climbing search, it is one of the most direct approaches, which is simply loop on the objective function graph to determine the local and global minimums and maximums. It continuously moves to the direction of increasing value and terminates once reaching a peak where no neighbor has higher value. It, neither maintains the search tree, nor does it look ahead searching for another local maximum. It reaches the solution rapidly with little memory, but it might have stuck in a local maximum/minimum. Hill-climbing search is not suitable when the graph has a flat local maximum, or a sequence of maximums that are hard to detect by the algorithm. So, the hill-climbing algorithm is guaranteed to be incomplete.

Simulated Annealing

Simulated annealing is the algorithm that combines hill-climbing with the random walk, to find the global maximum in a much more better way than simple hill-climbing, in both

efficiency and completeness. In metallurgy, annealing is the process used to temper or harden metals and glass by heating them to a high temperature and then gradually cooling them, thus allowing the material to reach a low energy crystalline state.

To explain simulated annealing, consider the gradient descent (i.e., minimizing cost) algorithm, and imagine the task of getting a ping-pong ball into the deepest crevice in a bumpy surface. If we just let the ball roll, it will come to rest at a local minimum. If we shake the surface, we can bounce the ball out of the local minimum. The trick is to shake just hard enough to bounce the ball out of local minima but not hard enough to dislodge it from the global minimum.

The simulated-annealing solution starts by shaking hard (i.e., at a high temperature) and then gradually reduce the intensity of the shaking (i.e., lower the temperature).

The simulated-annealing solution starts by picking a random move, if it improves the situation, it is always accepted. Otherwise, the algorithm accepts the move with some probability less than 1. The probability decreases exponentially with the “badness” of the move. The probability also decreases as the “temperature” T goes down: “bad” moves are more likely to be allowed at the start when T is high, and they become more unlikely as T decreases. If the schedule lowers T slowly enough, the algorithm will find a global optimum with probability approaching 1.

Simulated annealing was first used extensively to solve VLSI layout problems in the early 1980s. It has been applied widely to factory scheduling and other large-scale optimization problems [See Russell S. et al. (2010)].

Quantum Annealing

Quantum annealing is a quantum algorithm for finding the global minimum of a given objective function over a given set of candidate states, by a process using quantum fluctuations. Quantum annealing is used mainly for problems where the search space is discrete (e.g., combinatorial optimization problems) with many local minima, such as finding the ground state of a spin glass.

It was formulated in its present form by Kadowaki and Nishimori [See Kadowaki T. et al. (1998)]. Finilla had proposed a quantum annealing algorithm in a different form [See Finilla A. B. et al. (1994)].

Quantum annealing starts from a quantum-mechanical superposition of all candidate states with equal weights. Then the system evolves following the time-dependent Schrödinger equation, as a natural quantum-mechanical evolution of physical systems. The amplitudes of all candidate states keep on changing, realizing a quantum parallelism, according to the time-dependent strength of the transverse field, which causes quantum tunneling between states.

If the rate of change of the transverse field is slow enough, the system stays close to the ground state of the instantaneous Hamiltonian. If the rate of change of the transverse field is accelerated, the system may leave the ground state temporarily but produce a higher likelihood of concluding in the ground state of the final problem Hamiltonian. The solution for the original optimization problem is reached when the system has reached the suitable ground state of the classical Ising model [See Venegas-Andraca S. E. et al. (2018)].

Quantum annealing can be compared to simulated annealing, whose “temperature” parameter plays a similar role to QA's tunneling field strength. In simulated annealing, the temperature determines the probability of moving to a state of higher “energy” from a single

current state. In quantum annealing, the strength of transverse field determines the quantum-mechanical probability to change the amplitudes of all states in parallel.

6. CONCLUSION

Quantum computing has been developing its major steps during the last few years, from both perspective of building quantum computing machines, and of designing quantum programming languages. One of the important aspects is to distinguish between the general purpose, universal machine-like IBM's machine and the D-Wave's single purpose quantum annealing machine. Another important point is that, so far, no one knows, which quantum hardware model will win the race of beating the decoherence problem. Building quantum hardware is challenging, yet few years are separating us from the first quantum machine that will achieve the quantum supremacy.

The top most two competitors as per today are IBM's superconductors model with the Qiskit simulator, and Microsoft's topological qubit model with the Q# simulator. A valid suggestion for future work is to compare the results of running two implementations for the same algorithm one for each simulator.

REFERENCES

- Feynman R.P. (1982), Simulating Physics with Computers, International Journal of Theoretical Physics, 21: 467–488
- Benioff P. (1980), The computer as a physical system: A microscopic quantum mechanical Hamiltonian model of computers as represented by Turing machines, Journal of Statistical Physics. 22 (5): 563–591
- Manin, Yu. I. (1980), Computable and Noncomputable, Sov.Radio. pp. 13–15
- Deutsch D. (1985), Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer. Proceedings of the Royal Society of London A. 400 (1818): 97–117
- Yanofsky N. S., Mannucci M.A. (2013), Quantum Computing for Computer Scientist. Reprinted. Cambridge University Press (First published 2008)
- Kaye P., Laflamme R., Mosca M. (2010), An Introduction to Quantum Computing. Reprinted with corrections. Oxford University Press, 2010 (First published 2007)
- Feynman R. P. (1963), Feynman Lectures on Physics (3 Volume Set), Addison-Wesley
- Lloyd S. (1996), Universal quantum simulators, Science 273 (5278): 1073–8
- Kim M. (2017), Google's quantum computing plans threatened by IBM curveball, NewScientist article in the technology section, 20 October 2017, available online: <https://www.newscientist.com/article/2151032-googles-quantum-computing-plans-threatened-by-ibm-curveball/>
- Jackson J. (2014), IBM questions the performance of D-Wave's quantum computer, PCWorld news article, 04 February 2014, available online: <https://www.pcworld.com/article/2094380/ibm-questions-the-performance-of-dwaves-quantum-computer.html>

- Russell J. (2017), Quantum Bits: D-Wave and VW; Google Quantum Lab; IBM Expands Access, HPC Wire article, 21 March 2017, available online:
<https://www.hpcwire.com/2017/03/21/quantum-bits-d-wave-vw-google-quantum-lab-ibm-expands-access/>
- Vincent J. (2016), Biggest ever quantum chip announced, but scientists aren't buying it, The Verge article, 28 September 2016, available online:
<https://www.theverge.com/2016/9/28/13057414/quantum-computer-d-wave-2000-qubit-chip>
- Neven H. (2015), When can Quantum Annealing win?, Google AI Blog article, 8 December 2015, available online:
<https://ai.googleblog.com/2015/12/when-can-quantum-annealing-win.html>
- Marchenkova A. (2016), What's the difference between quantum annealing and universal gate quantum computers?, Medium article, 27 February 2016, available online:
<https://medium.com/quantum-bits/what-s-the-difference-between-quantum-annealing-and-universal-gate-quantum-computers-c5e5099175a1>
- Heath N. (2018), D-Wave quantum computers: Cheat sheet, TechRepublic article, 3 July 2018, available online:
<https://www.techrepublic.com/article/d-wave-quantum-computers-the-smart-persons-guide/>
- [IBM Quantum Computer] [Infograph] (n.d.), retrieved from <https://www.research.ibm.com/ibm-q/learn/what-is-ibm-q>, the exact URL: <https://www.research.ibm.com/ibm-q/learn/what-is-ibm-q/images/infographic-inside.jpg>
- Willet R.L. (2013), Magnetic field-tuned Aharonov–Bohm oscillations and evidence for non-Abelian anyons, at $\nu = 5/2$. arXiv:1301.2639
- Von Keyserling C., Simon S. H., Bernd R. (2015), Enhanced Bulk-Edge Coulomb Coupling in Fractional Fabry-Perot Interferometers, Physical Review Letters. 115
- Russell S., Norving P. (2010), Artificial Intelligence a Modern Approach. Third edition, Prentice Hall (First Published 1995)
- Kadowaki T., Nishimori H. (1998), Quantum annealing in the transverse Ising model, Phys. Rev. E 58, 5355
- Finilla A. B., Gomez M. A., Sebenik C., Doll J. D. (1994), Quantum annealing: A new method for minimizing multidimensional functions, Chem. Phys. Lett. 219, 343
- Venegas-Andraca S. E., Cruz-Santos W., McGeoch C., Lanzagorta M. (2018), A cross-disciplinary introduction to quantum annealing-based algorithms, Contemporary Physics Vol. 59, Issue 02, pp. 174–196

Improving Storage Capacity of QR Code: A Survey

Islam M. El-Sheref¹

Ahmed H. Asad¹

Fatma A. El-Licy¹

Abstract

QR codes have become widely popular along with its useful applications. Its specifications offer many more advantages than that of the one-dimensional code. In particular, the advantages of the colored QR Code, encourage its enhancement to improve the capacity and reduced the effects of the image distortion. The paper presents the current research advances for improving quick response code storage capacity.

Keywords: 2-D barcodes, QR Codes - Colored QR Codes - Error Correcting Codes - Image multiplexing - Data Capacity Enhancement.

1. Introduction

The Quick Response (QR) code is created by the Toyota subsidiary Denso Wave in 1994 and was initially used for tracking inventory in vehicle parts manufacturing [see Ahlawat et al. (2017)]. A QR code is a type of matrix or two-dimensional barcode that stores up to 4296 characters of information and it is designed to be read by smartphones [see Tiwari (2016)]. QR stands for “Quick Response” indicating that the code contents should be decoded very quickly at high speed. A QR code consists of black modules arranged in a square pattern on a white background. The information encoded may be text, a URL or other data [see Singh and Singh (2016)]. The idea behind the development of the QR code is the limitation of the standard Universal Product Code (UPC) barcode information capacity (that holds, only, 20 characters) [see Kinjal et al. (2014)]. QR codes system became popular due to its fast readability and greater storage capacity [see Rizqi et al. (2018), Bhardwaj et al. (2016), Goyal et al. (2016), Čović et al. (2016)].

1.1. Types of 2D Code

There are many types of 2D Code each of which has unique structure. Figure 1.A, shows the structure of the QR code, while figure 1.B, depicts the structure of other types of two-dimension codes. Yet, QR codes are utilized most often for its higher data storage capacity.

¹ Computer Sciences department, Institute of Statistical Studies and Research, Cairo University

QR codes have become widely popular along with its useful applications. Its specifications offer many more advantages than that of the one-dimensional code. The Advantages of QR Code include:

- Reduced space.
- Durability against soil and damage.
- High data capacity.
- More supported languages.
- Supporting of 360-degree reading.

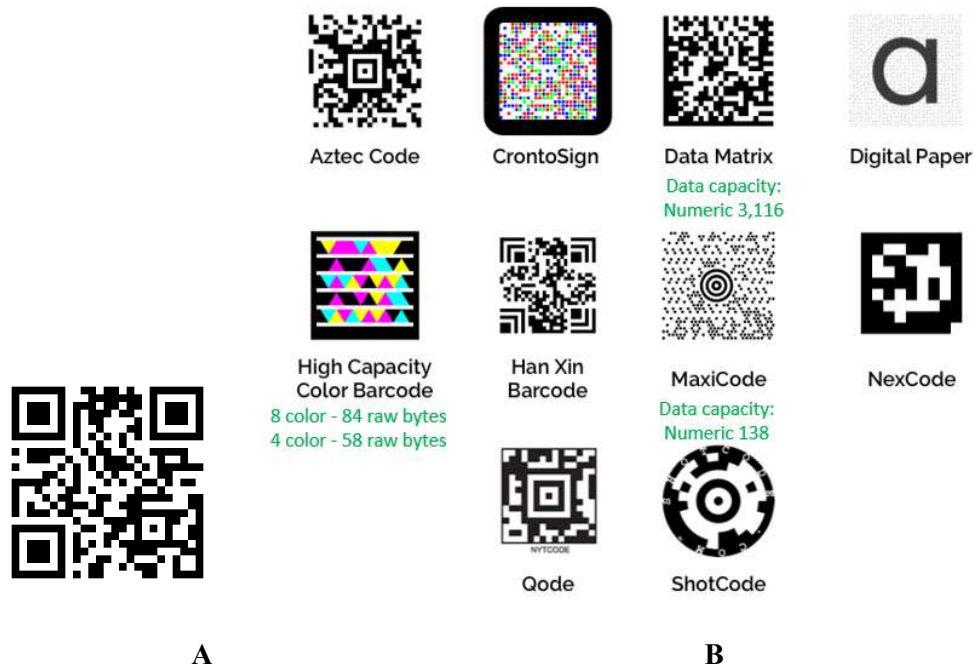


Figure 1: A. QR Code, B. and other type of 2D code

1.2. Application on QR Code

Based on the flexibility of QR code, there are many applications on QR Code, including:

- Mobile operating systems.
- Virtual stores.
- Payment.
- Website login, Wi-Fi network login and TOTP.
- Robot Navigation.
- Education.

1.3. QR Code Structure

The image that represents a QR code composed of three main constructions, namely: function pattern, encoding region and quiet zone.

1.3.1. Function patterns

Must be placed in specific areas of the QR code to ensure that it can be correctly identified by QR code scanners. The function patterns are:

1. Finder pattern: are the special position-detection patterns located in three corners (upper left, upper right, and lower left) of each symbol.
2. Separators: They are the one-module wide areas of whitespace between each finder pattern and encoding region.
3. Timing patterns: There are 2 timing patterns (horizontal timing pattern and vertical timing pattern), they are consisting of alternating dark and light modules. These patterns are helpful in determining the symbol density, module coordinates and version information area.
4. Alignment patterns: They are a construction of three modules: A 5×5 dark modules, surrounding a 3×3 light modules and a single dark module in the center. QR codes, that are version 2 and higher, must have alignment patterns and the number of alignment patterns depends on the symbol version.

1.3.2. Encoding region

Encoding region contains the information about the coded data, as shown in Figure 2:

- Version information
- Format information: contains format error correction, mask pattern and error correction level (EC Levels) as illustrated in Figure 3, and detailed in Section 1.4.
- Error correction codewords: contains the encoded data after adding error correction codeword

1.3.3. Quiet Zone

It is a 4-module wide area containing no data.

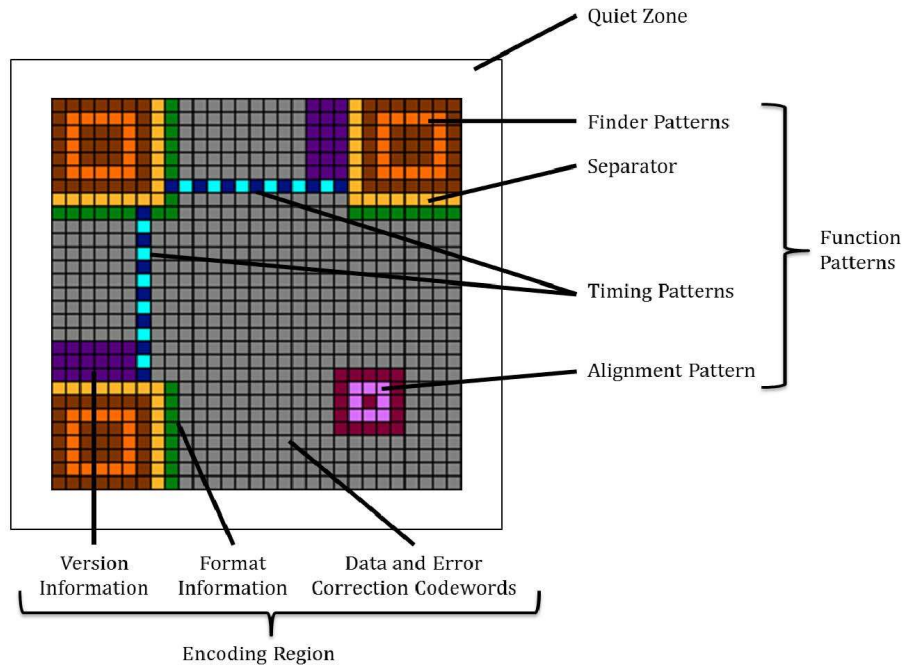


Figure 2: Structure of QR Code

1.4. Format Information

The left area in Figure 3, illustrates the left half region of the QR code structure. The format information is coded in three location:

- The vertical bar parallel to the upper left finder.
- The horizontal bar below the left finder pattern.
- The vertical bar parallel to the lower left finder pattern.

Error correction level:

Codewords are 8 bits long and use the Reed–Solomon error correction algorithm with four error correction levels. The higher the error correction level, the less storage capacity. Table 1, lists the approximate error correction capability at each of the four levels, Figure 3. For example, at the error correction Level M (Medium), 15% of codewords can be restored.

Table 1: Error correction levels and percentage of correction

S No.	Error-Correction Level	Approximate Amount of Correction
1.	L	7%
2.	M	15%
3.	Q	25%
4.	H	30%

Mask pattern:

It is used to break up patterns in the data area that might confuse a scanner, such as large blank areas or misleading features that look like the locator marks. The mask patterns are defined on a grid that is repeated as necessary to cover the whole symbol. Modules corresponding to the dark areas of the mask are inverted to white. Figure 3, illustrate the masking patterns.

1.5 QR Code Versions and Storage Capacity

The symbol versions of the QR Code may range from version 1 to version 40, each of which has a different module configuration or number of modules. The module refers to the black and white dots that make up QR code [see Ahlawat et al. (2017), Tiwari (2016), Singh and Singh (2016), Kinjal et al. (2014)]. Table 2, illustrates the versions of QR Code.

1.6. QR Code Working Overview

The QR code system consists of a QR code encoder and decoder. The encoder is responsible for encoding data and generation the QR Code, while the decoder decodes the data from the QR code. Figure 4, shows an overview of the QR code working. The plain text, URL, or other data are given to the QR code encoder, generate the required QR code. A scanner (decoder) however, is utilized to decode QR code back to its original data.

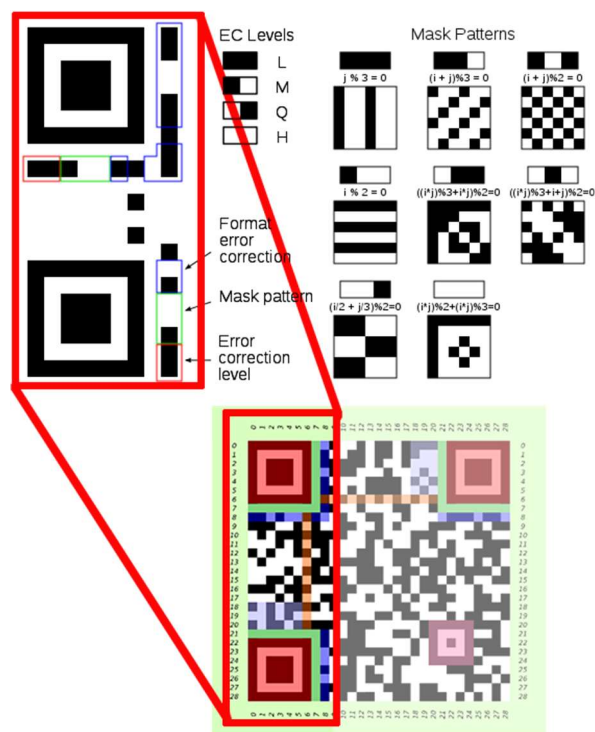


Figure 3: Format information

Table 2: Data Capacity of QR Code version 40

Version	Modules	ECC Level	Data Bits (mixed)	Numeric	Alpha-numeric	Binary	Kanji
40	177x177	L	23,648	7,089	4,296	2,953	1,817
		M	18,672	5,596	3,391	2,331	1,435
		Q	13,328	3,993	2,420	1,663	1,024
		H	10,208	3,057	1,852	1,273	784

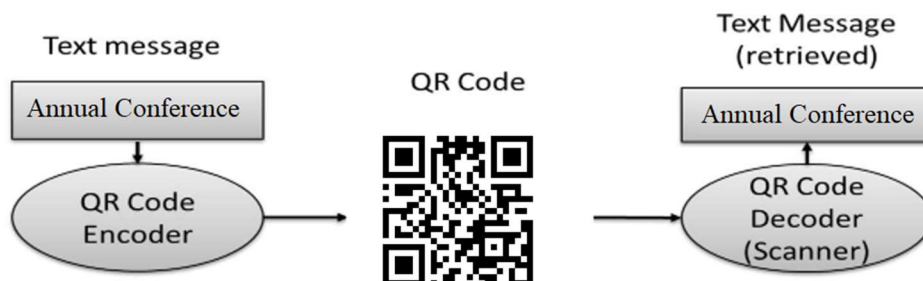


Figure 4: Working (overview) of QR Code

2. Extending QR Code Capacity

The structural flexibility of QR code architecture opens many more possibilities to stretch the limits of data storage capacity. The following are some of these possibilities:

- Data hiding techniques
- Multiplexing techniques
- Using data compression techniques.
- Using colored QR codes.

3. Improving QR Capabilities

There are many applications that focus in improving QR code data capacity and its tolerance to distortion. That is influenced by QR code structural flexibility. Some of these techniques focus on increasing data capacity, including data hiding techniques and data compression techniques [see Gupta et al. (2018)]. Data hiding techniques distort the QR code during the hiding process, which can be resolved by using additional algorithms for the correction of image distortions. The amount of increasing in data capacity, due to data hiding technique, however, is limited [see Paul et al. (2018)]. Another way to increase the data capacity is to compress the data before generating a QR code, results, however, shows that data compression technique offers compression up to 52% of the original size [see Abas et al. (2017)].

3.1. Coloured QR Codes "CQR Code-5":

Vizcarra [see Melgar et al. (2012)] purposed a colored QR code structure. It is designed to employ 5 different RGB colors (red, green, blue, black and white), as shown in Figure 5, which enables twice as much as that of the traditional binary QR codes. Each information module represents 2 bits, while the traditional utilize 1 bit for each module. Black modules are used only for alignment purposes. The “red, green, blue and white” colors are chosen because of their maximum equidistance on the RGB color space, as shown in Figure 6.

CQR Code-5 has the following advantages:

- a) Increasing of data storage capacity, Two times the traditional QR Code.
- b) Applying the same localization and geometry correction algorithms for monochrome barcodes on the grayscale image.

CQR Code-5 Disadvantage is that it ignores the color distortions caused by printer while printing QR code or by camera sensor and surrounding lighting in capture process.

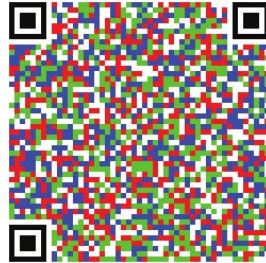


Figure 5: Example of a CQR code

Table 3: Module represent in CQR code-5

Color	Module represent
Red	00
Green	01
Blue	10
White	11

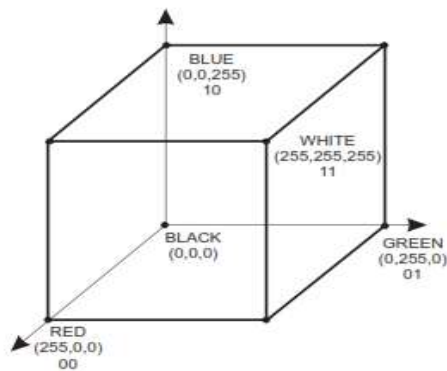


Figure 6: The equidistance on RGB color space

3.2. CQR Code-9:

Melgar [see Melgar et al. (2016)] proposed CQR Code-9 Figure 7. It can store up to 2,024 information bits which is twice that of CQR Code-5 and 4 times that of classic QR code. Each information module represents 3 bits. The color code is shown in Table 4.

Advantages of CQR Code-9:

- a) Increasing of data storage capacity, Three times the traditional QR Code.
- b) applying the same localization and geometry correction algorithms for monochrome barcodes on the grayscale image.

Disadvantage of CQR Code-9, though, is that it ignores the color distortions caused by printer while printing QR code or by camera sensor and surrounding lighting during the capture process.

Table 4: Module represent in CQR code-9

Color	Module represent
Red	000
Green	001
Blue	010
Cyan	011
Magenta	100
Yellow	101
White	110
Gray	111



Figure 7: Example of CQR Code-9.

3.3. Per-Colorant-Channel Color Barcodes:

Blasinski [see Blasinski (2013)], exploit the spectral diversity afforded by the cyan (C), magenta (M), and yellow (Y) print colorant channels commonly used for color printing and the complementary red (R), green (G), and blue (B) channels, respectively, used for capturing color images. Specifically, exploiting this spectral diversity to realize a three-fold increase in the data rate by encoding independent data in the C, M, and Y print colorant channels and decoding the data from the complementary R, G, and B channels captured via a mobile phone camera. Developing interference cancellation algorithm to mitigate the effect of cross-channel interference among the print colorant. To estimate the model parameters required for cross-channel interference cancellation, He proposed two alternative methodologies "Pilot block" approach and an "Expectation maximization (EM)-type" approach. Data was encoded using three layers of C, M, and Y which are printed in overlay to generate the color barcode Figure 8, Data encoded to each layer independently. Data recovery of the captured QR code is first localized and separated into R, G, and B color channels, typically used in capture devices then using color interface cancellation to overcome the interference between the two-color system then extract the corresponding CMY layer to decode data as shown in Figure 9.

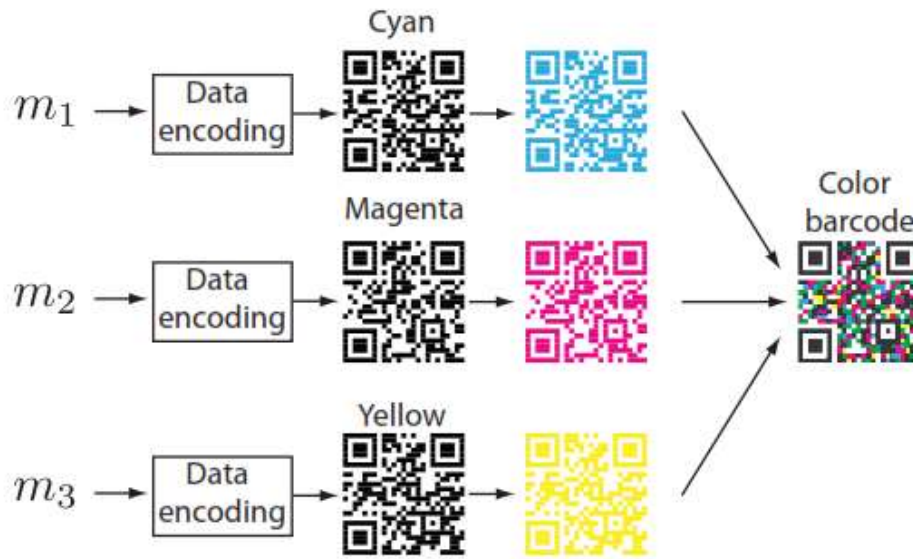


Figure 8: Per-channel-based data encoding.

Advantages of Per-Colorant-Channel:

- a) Increasing of data storage capacity, Three times the traditional QR Code.

- b) Using the same localization and geometry correction algorithms for monochrome barcodes on the grayscale image.
- c) Avoiding interference between CMYK and RGB color systems.

Disadvantage of Per-Colorant-Channel is that it Ignores the color distortions caused by camera sensor and surrounding lighting during the capture process.

3.4. Use Multiplexing to Increase Information in QR Code:

Vongpradhip [see Vongpradhip (2013)] purpose Multiplexing QR Codes, in encoding the original data is divided into smaller parts; each part will form QR code pattern in its standard form. After that, each QR code pattern is multiplexed and represented as module in the final QR code with black and white special symbols as shown in Figure 10.

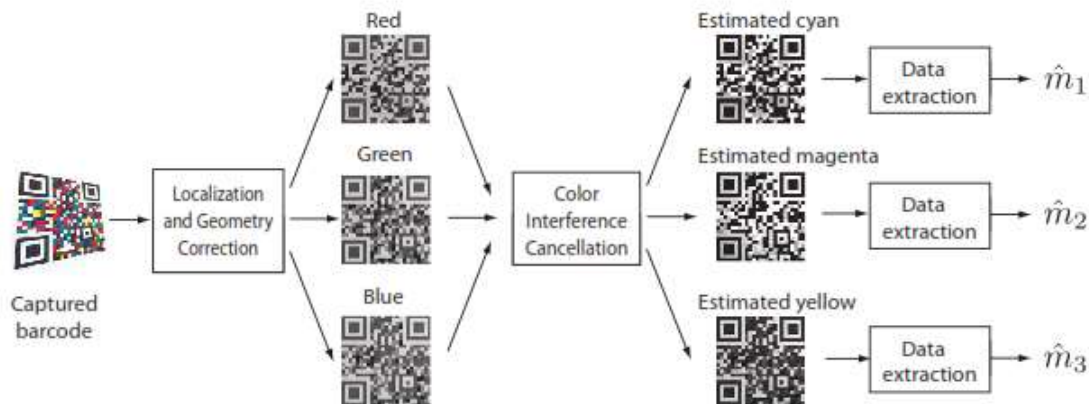


Figure 9: Data recovery from captured QR code.

Decoding: QR code with special symbols (that was multiplexed) is decoded to give back the number of QR code patterns that was multiplexed. These QR code patterns can be read by the general QR code reader and the data can be concatenated back to form its original information. The number of required symbols is defined by $2n$ where n is the number of QR code patterns. As an example, the sentence “Sartid Vongpradhip Chulalongkorn U.” is divided into three parts:

- a) Sartid,
- b) Vongpradhip and
- c) Chulalongkorn U.

A QR code pattern is then generated for each part using the special symbols shown in Table 5, for multiplexing three QR code patterns in one. Then at decoding, demultiplex them “using pattern recognition” into three traditional binary QR codes.

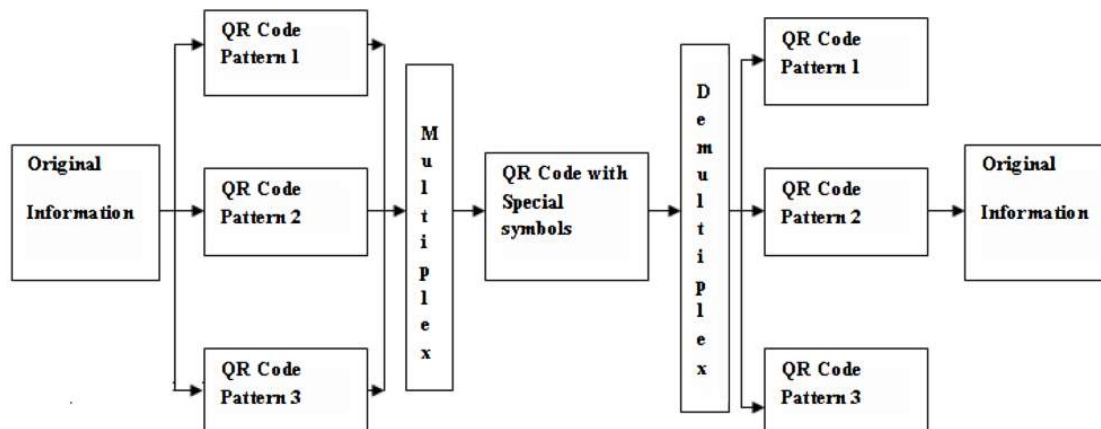


Figure 10: Overview of multiplexing and demultiplexing approach.

The main advantage of Multiplexing is the increase of data storage capacity.

Disadvantages of Multiplexing are:

- Increasing of data storage capacity is limited by the number of available shapes that a pattern recognition algorithm can recognize.
- A pattern recognition algorithm is required to detect these black and white symbols which make decoding procedure very much different from that of the basic QR code.

Table 5: Module represent in Multiplexed QR code

Module in each pattern	Symbols
000	\
001	/
010	√
011	∧
100	>
101	<
110	└
111	┘

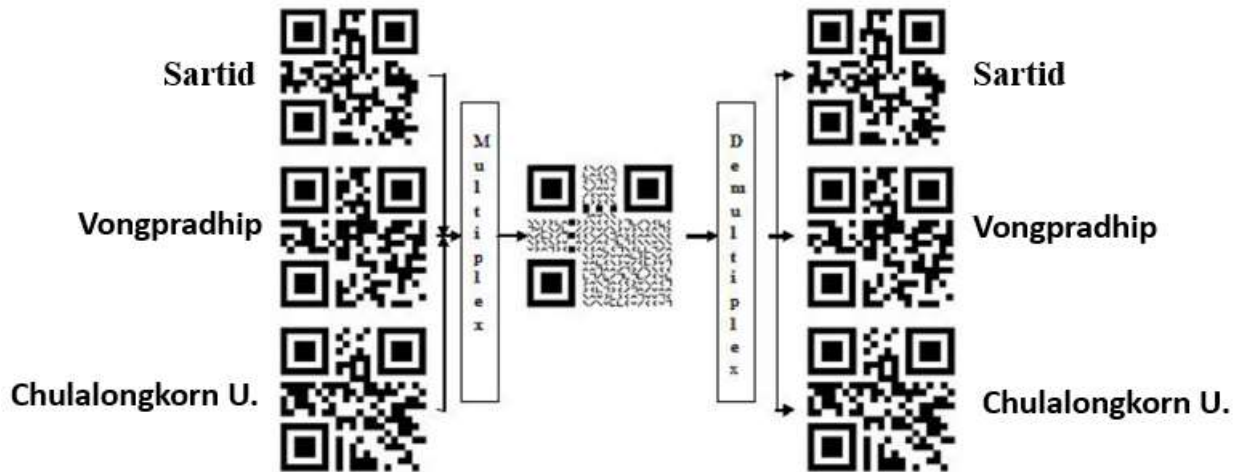


Figure 11: Experiment with 8 special symbols.

3.5. Extending the Storage Capacity for Faster QR-Code:

Gupta [see Gupta et al. (2018)] propose a new approach for extending the storage and faster QR decoding. Encoding: Convert part of data to normal QR code to generate 1st QR code, in which the black bits are one third instead of full of 1st QR code. Convert another part of data to normal QR code to generate 2nd QR code, in which black bits are instead of full of 2nd QR code, then rotate 2nd QR code in 90 degrees and place it on 1st QR code Figure 12. Then place a black bit in top-left most corner, as illustrated in Figure 13. The author performs the decoding by utilizing bar code decoding system, in which black-colored blocks with length less than the module size is ignored. The horizontal and vertical regions are scanned and the black modules with size less than one third that of the standard are ignored, as shown in Figures 15 and 16.

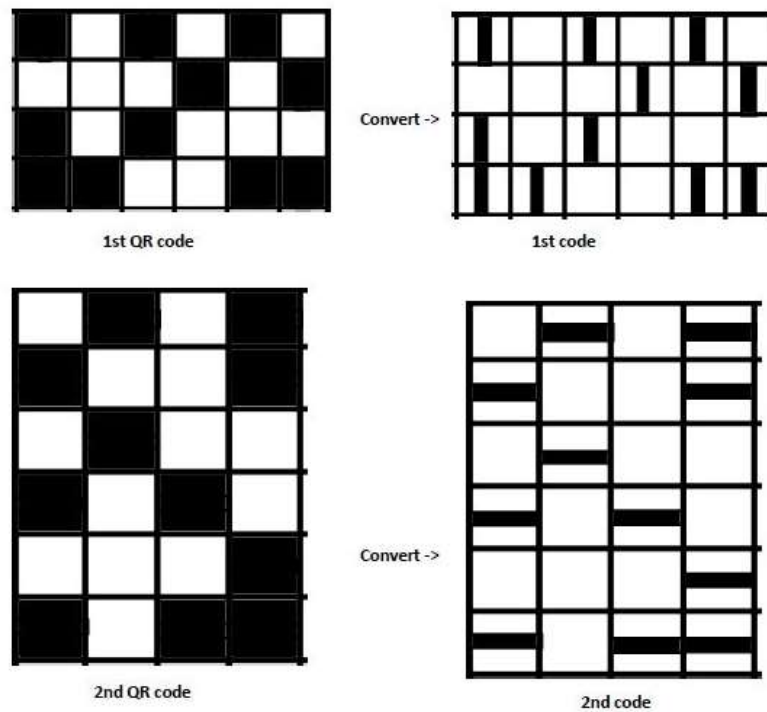


Figure 12: Decoding and rotating QR code.

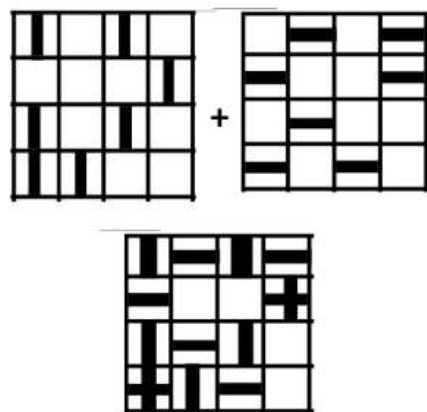


Figure 13: Placing 1st QR code and 2nd QR code.



Figure 14: Encoding Experimental result “Hello World”

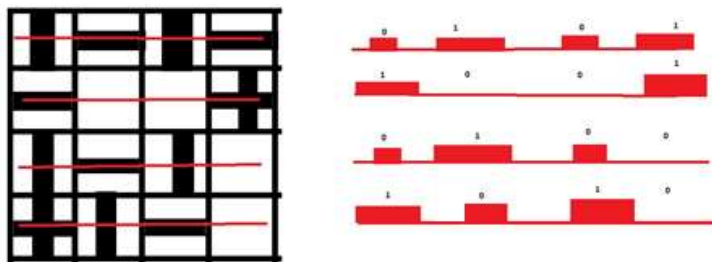


Figure 15: Decoding vertical scanning

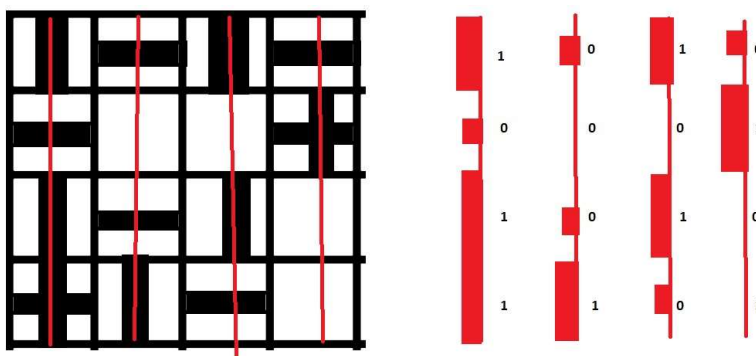


Figure 16: Decoding horizontal scanning

Advantages of Gupta approach are:

- a) Increase of data storage capacity up to 2 times that of traditional binary QR Codes.
- b) Faster decoding than classic QR code.

Disadvantage of Gupta approach:

- a) Waste some space and if data when data gets bigger then waste of space gets more visible. As shown in Figure 17.
- b) It will not be recognized if data is large and its image is presented in a small area.
- c) More vulnerable to light glare and illumination issue.
- d) Is require additional step (barcode decoding).

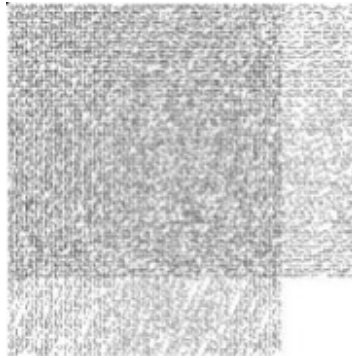


Figure 17: Waste of space in right bottom square corner.

3.6. Use Multiplexing with Color Coding:

Galiyawala [see Galiyawala and Kinjal Pandya (2014)] intended MUX method in which several QR codes are mixed to generate QR code, this technique offers increasing of the data storage capacity up to 24 times compared to a standard QR code of the same size. Encoding: Original information is divided into "n" parts, which are used to generate "n" individual QR codes, so it needs $2n$ distinct encoding color as shown in Table 6.

Example for encoding the three QR code we need $2^3 = 8$ distinct color, if the first modules in the three QR code are "white, white, black" = "0,0,1" it will be multiplexed into blue color according to Table 7.

Decoding: the color value is demultiplexed using the table to get the three-separate classic QR codes.

Advantages of Multiplexing with Color Coding:

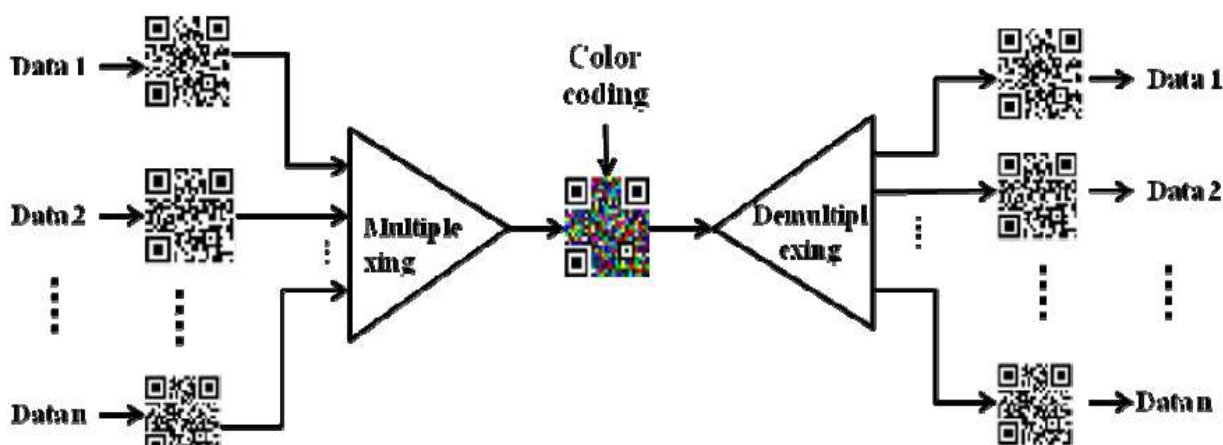
- a) Increasing of data storage capacity up to 24 times traditional binary QR Codes.
- b) Using the same localization and geometry correction algorithms for monochrome barcodes on the grayscale image.

Disadvantage of Multiplexing with Color Coding:

- a) Ignoring the color distortions caused by printer while printing QR code or by camera sensor and surrounding lighting during the capture process.
- b) The processing time for multiplexing and especially that for demultiplexing increased exponentially when merging more than 9 images as shown in Table 8.

Table 6: Distinct color requirements for QR code multiplexing

Number of QR codes for multiplexing (n)	Required distinct colors (2^n)
2	4
3	8
4	16
5	32
6	64
7	128
8	256
...	...
14	16384

**Figure 18:** Block diagram of multiplexing with color coding technique.

Advantages of Multiplexing with Color Coding:

- a. Increasing of data storage capacity up to 24 times traditional binary QR Codes.
- b. Using the same localization and geometry correction algorithms for monochrome barcodes on the grayscale image.

Disadvantage of Multiplexing with Color Coding are:

- Ignoring the color distortions caused by printer while printing QR code or by camera sensor and surrounding lighting during the capture process.
- The processing time for multiplexing and especially that for demultiplexing increased exponentially when merging more than 9 images as shown in Table 8.

Table 7: Multiplexing and color-coding table for three QR codes

QR code 1	QR code 2	QR code 3	Index	Red	Green	Blue	Color
0	0	0	0	0	0	0	Black
0	0	1	1	0	0	1	Blue
0	1	0	2	0	1	0	Green
0	1	1	3	1	0	0	Red
1	0	0	4	0	1	1	Cyan
1	0	1	5	1	0	1	Magenta
1	1	0	6	1	1	0	Yellow
1	1	1	7	1	1	1	White

Table 8: Observation table for processing time

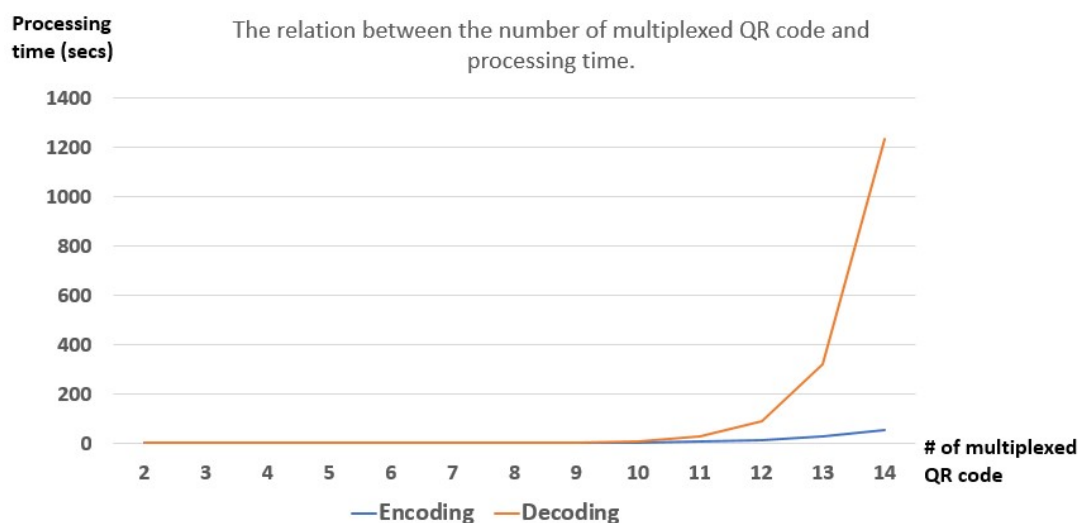


Figure 19: Number of multiplexed QR Code versus its processing time.

Sr. No.	Number of QR codes multiplexed	Assigned distinct colors (2^n)	Processing time (seconds)	
			<i>During Encoding</i>	<i>During Decoding</i>
1	2	4	2.922	2.892
2	3	8	3.194	2.988
3	4	16	3.364	2.982
4	5	32	3.131	3.084
5	6	64	3.297	3.381
6	7	128	3.489	3.624
7	8	256	3.911	4.359
8	9	512	4.811	6.128
9	10	1024	6.229	11.406
10	11	2048	9.453	28.398
11	12	4096	15.787	91.393
12	13	8192	27.940	323.198
13	14	16384	53.157	1236.105

4. Tools of QR Code

The software industry provided varieties of tools to support the recognition and creation of QR for different purposes and plate forms. The development of QR is promising for its advantages and storage capabilities.

4.1. QR Code Readers

Barcode Scanner/Reader Scan barcodes on products, look up prices and reviews. Also scan QR codes containing URLs. Share contacts, apps, and bookmarks via QR code. Here are some of the current QR readers:

I-nigma: claims to be the most widely used mobile barcode reader in the world. Decode mobile barcodes and connect directly to the mobile Internet. Features include customizable sound, automatic saving, and automatic launching of URLs. Save scanned contact information into contacts. Available for iPhone, Android, Blackberry.

QR App: Designed to be fast and simple. Display history of scanned QR codes. Add contacts directly to address book. Scan mode automatically detects code, takes photo and decodes. Available on iPhone.

QR Droid: Scan in both landscape and portrait modes, or even scan mirrored images. Decode saved images or URL addresses of a QR code. Generate a QR code to share contacts, bookmarks, apps, geolocations, and more. Available on Android.

QR Reader for iPhone: QR Reader for iPhone scans quickly, with a range of usability and sharing options. Text and URLs open within the app. Launch additional apps to open information in details view. There's also a built-in web browser. Tap 'Share' brings up email, Facebook, Twitter and copy options. Post on Facebook and Twitter without leaving the app.

4.2 QR Code Generators

The following are some of the current QR code generator:

Kaywa: is a site created by Datamatrix, a pioneer of 2D codes. Use it to create QR codes that link to URLs, text, phone numbers, and SMS.

QR Code and 2D Code Generator: generates QR Code, Micro QR Code, Data Matrix and Aztec Code. Additional features include URL shortening and customizing the color and format of your codes.

Google's Visualization API: is a powerful API capable of generating interactive charts, but it can also be used to create QR codes.

5. Discussion:

This paper introduced the quick response (QR) code and its applications with the possibilities of its utilization in different areas that is yet to be explored. The technology has a firm ground for research aspects. More and more experiments are done with QR Codes in different aspects like enhancing the storage capacity, better recognition, reducing redundancy in order to save space, possibility of encoding different kind of data like audio, etc. As QR Code provides the structural flexibility, it opens up the huge platform for researchers to explore the possibilities to enhance the performance of QR Code and increase storage capacity.

References

- Abas A., Yusof Y. and Kabir F., 2017, "Expanding the data capacity of QR codes using multiple compression algorithms and base64 (encode/decode)", Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 9, no. 2-2, pp. 41-47.

- Bhardwaj N., Kumar R., Verma R., Jindal A. and Bhondekar A. P., 2016, "Decoding algorithm for color QR code: A mobile scanner application", International Conference on Recent Trends in Information Technology (ICRTIT), India, pp. 1-6.
- Goyal S., Yadav S. and Mathuria M., 2016, "Exploring concept of QR code and its benefits in digital education system", International Conference on Advances in Computing, Communications and Informatics (ICACCI), India, vol. 3, issue 5, pp. 452-456.
- Gupta K. D., Ahsan M. and Andrei S., 2018, "Extending the storage capacity and noise reduction of a faster QR-code", Broad Research in Artificial Intelligence and Neuroscience (BRAIN), vol. 9, issue 1, pp. 59-71.
- Henryk Blasinski, April 2013, "Per-Colorant-Channel Color Barcodes for Mobile Applications: An Interference Cancellation Framework", vol. 22, no. 4, pp. 1498-1511.
- Hiren J. Galiyawala and Kinjal H. Pandya, 2014, "To Increase Data Capacity of QR Code Using Multiplexing with Color Coding: An example of Embedding Speech Signal in QR Code", IEEE India Conference (INDICON), pp. 1-6.
- Kinjal H. Pandya and Hiren J. Galiyawala, 2014, "A survey on QR codes: in context of research and application", International Journal of Emerging Technology and Advanced Engineering (IJETA), vol. 4, issue 3, pp. 258-262.
- Kishor Datta Gupta, Md Manjurul Ahsan and Stefan Andrei, 2018, "Extending the Storage Capacity and Noise Reduction of a Faster QR-Code", Lamar University, Beaumont, USA pp. 59-71.
- Max E. Vizcarra Melgar, Alexandre Zaghetto, Bruno Macchiavello, Anderson C. A. Nascimento, 2012, "CQR CODES: COLORED QUICK-RESPONSE CODES", IEEE Second International Conference on Consumer Electronics - Berlin (ICCE-Berlin), pp. 3321-3325.
- Max E. Vizcarra Melgar, Mylène C. Q. Farias, Flávio de Barros Vidal, Alexandre Zaghetto, 2016, "A High Density Colored 2D-Barcode: CQR Code-9", SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 329-334.

- Remya Paul, 2018, "A review on stegnography in QR codes", International Research Journal of Engineering and Technology (IRJET), vol. 5, issue 5, pp. 4294-4296.
- Rizqi R.I., Rohama N. A. and dr. Nimkerdphol K., 2018, "Inventory management system using QR code on android a case Study in computer engineering department", Journal of Electrical Engineering and Computer Sciences (JEECS), vol. 3 no. 1, pp. 381-388.
- Sartid Vongpradhip, 2013, "Use Multiplexing to Increase Information in QR Code", Colombo, Sri Lanka, pp. 361-364.
- Seema Ahlawat, 2017, "A review on QR codes: colored and image embedded" International Journal of Advanced Research in Computer Science (IJARCS), vol. 8, no. 5, pp. 410-413.
- Singh A. and Singh P., 2016, "A review: QR codes and its image pre-processing method", International Journal of Science, Engineering and Technology Research (IJSETR), vol. 5, no. 6, pp.1955-1960.
- Sumit Tiwari, 2016, "An introduction to QR code technology", IEEE International Conference on Information Technology (ICIT), pp. 39-44.
- Zlatko Čović, Ürmös Viktor, János Simon, Dalibor Dobrilović, Željko Stojanov, 2016, "Usage of QR Codes in Web Based System for the Electronic Market Research", 14th IEEE International Symposium on Intelligent Systems and Informatics (SISY), Serbia, pp. 187-192.

Record Linkage Techniques for Handling the Volume Property of Big Data

Randa M. Abd El-Ghafar¹ Ali H. El-Bastawissy² Eman S. Nasr³ Mervat H. Gheith¹

Abstract: Record Linkage aims to find records in a dataset that represent the same real-world entity across many different data sources. It is a crucial task for data quality. With the evolution of Big Data, new difficulties appeared to deal mainly with the 5Vs of Big Data properties; i.e. Volume, Variety, Velocity, Value, and Veracity. Therefore, Record Linkage in Big Data is more challenging. This paper investigates ways to apply Record Linkage techniques that handle the Volume property of Big Data and conducts a comparative study between the applied Record linkage techniques of big data. Our investigation revealed four major issues. First, the techniques used to resolve the Volume property of Big Data mainly depend on partitioning the data into a number of blocks. The processing of those blocks is parallelly distributed among many executors. Second, MapReduce is a famous programming model that is designed for parallel processing of Big Data. Third, Spark is an opensource cluster-computing framework used for large-scale data processing. Spark gains a lot of popularity due to its superiority to process the data in memory, which make it faster more than 100 X than MapReduce. Fourth, a blocking key is usually used for partitioning the big dataset into smaller blocks; it is often created by the concatenation of the prefixes of chosen attributes. Partitioning using a blocking key may lead to unbalancing blocks, which is known as data skew, where data is not evenly distributed among blocks. An uneven distribution of data degrades the performance of the overall execution of the MapReduce model. Fourth, to the best of our knowledge, a small number of studies has been done so far to balance the load between data blocks in a MapReduce framework. Hence, more work should be dedicated to balancing the load between the distributed blocks.

Keywords: Big Data; Big Data Integration; blocking; entity matching; entity resolution; Hadoop; machine learning; MapReduce; Record Linkage.

1. INTRODUCTION

The term Big in Big Data does not only refer to the size of the data, but it also indicates many characteristics known as the five dimensions of Big Data, or 5Vs of Big Data, as shown in Fig. 1. Those 5Vs are; Volume which refers to "Big" in the term Big Data; Variety which refers to Data that come from a variety of sources and has many formats (structured as traditional databases, semi-structured as XML files or non-structured as Images); Velocity which refers to how quickly the data could arrive, be stored, and retrieved; Veracity which represents the reliability of data; and Value which is the high value of data that could be obtained by analyzing a huge amount of data. Actually, the majority of the data (about 80%) is unstructured data and this explains why a business is mainly concerned with managing unstructured data (Dhavapriya & Yasodha, 2016).

¹Institute of Statistical Studies and Research (ISSR), Cairo University, Egypt

²Independent Researcher

³Faculty of Computer Science, Modern Sciences and Arts University, Egypt

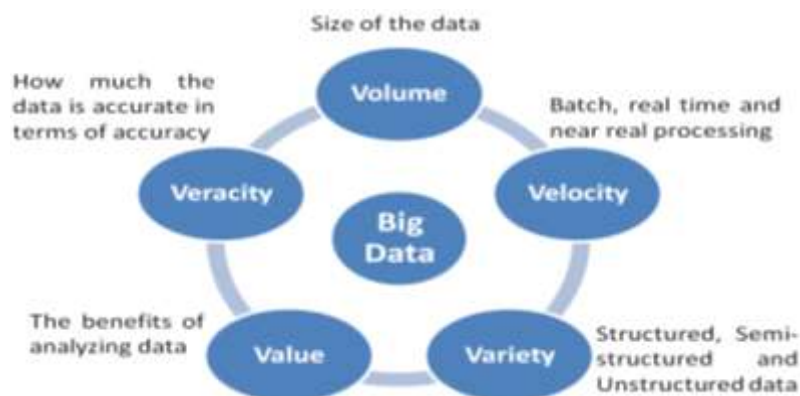


Fig. 1. Five dimensions of Big Data.

Data Integration is the process of combining data from many different sources and providing a user with one unified view of these data. Big Data Integration differs from traditional Data Integration in the three basic dimensions of Big Data known as 3Vs (Volume, Variety, Velocity), as they make the process of Big Data Integration more challenging and complicated. For the Volume, the number of data sources has grown to be millions for a single domain (Dong & Srivastava, Big Data Integration, 2015). For the Velocity, many of the data sources are very dynamic and rapidly exploded as a direct result of the rate at which data is moving and being collected from heterogeneous data sources (Dong & Srivastava, Big Data Integration, 2015). For Variety, Big Data projects include data sources from different domains, which are naturally diverse as they refer to different types of entities and relationships; these different types of sources are in need to be integrated into a unified view of data (Dong & Srivastava, Big Data Integration, 2015).

The architecture of Data Integration is composed of three steps, Schema alignment, Record Linkage and data fusion. The three steps aim to address the challenges of integrating data from multiple sources as semantic ambiguity, instance representation ambiguity, and data inconsistency (Dong & Srivastava, Big Data Integration, 2015) as shown in Fig. 2.

Schema alignment addresses the challenges of semantic ambiguity and aims to recognize which attributes have the same meaning and which does not (Dong & Srivastava, Big Data Integration, 2015). Record Linkage addresses the challenges of the ambiguity of instance representation and aims to recognize which elements represent the same real-world entity and which are not (Dong & Srivastava, Big Data Integration, 2015). Data fusion addresses the challenges of data quality and aims to know which values to use in case of conflict attributes value arises from many sources (Dong & Srivastava, Big Data Integration, 2015).

This paper investigates the Record linkage techniques to handle the volume challenge of big data. Adding the volume dimension to the traditional Record Linkage makes the Record Linkage more complicating and challenging as it compares a pair of records using one or more similarity techniques, which make Record Linkage more expensive task that could take days to be accomplished.

The rest of this paper is organized as follows. Section 2 presents a state of art of Big Data Integration Techniques. Section 3 discusses Record Linkage techniques that handle

the volume property of Big Data. Finally, the conclusion and future work are given in section IV.

2. STATE OF ART OF BIG DATA INTEGRATION TECHNIQUES

The importance of Big Data Integration resulted in increasing amounts of researches in the fields of schema mapping, Record Linkage and data fusion of Big Data Integration over the past few years to deal with the challenges associated with them. Table I shows a summary of these techniques (Dong & Srivastava, Big data integration, 2013).

2.1 Record Linkage Definition and Importance

Preparing and cleaning any dataset for analysis is very important step because of the concept of garbage in garbage out. It is a costly process, error pruning and time-consuming. Actually, it takes about 80% of the whole time spent on Preparing and cleaning any dataset for analysis is very important step because of the concept of garbage in garbage out. It is a costly process, error pruning and time-consuming. Actually, it takes about 80% of the whole time spent on analytics. According to Gartner, \$44 billion total in 2014 alone was invested for successfully preparing data obtained from many sources for use in data analysis (Castanedo, 2015).

Record Linkage, Entity Resolution or Entity Matching is a crucial step in data cleaning which aims to find records (i.e., database tuples) that refer to the same real-world entity (Getoor & Machanavajhala, 2012), (Kong, Gao, Xu, Qian, & Zhou, 2016). Record Linkage is an expensive process that could take many hours or even days. The situation become more complicated especially for large datasets as it compares a pair of records using one or more similarity measures (Köpcke, Thor, & Rahm, 2010).



Fig. 2. Traditional Data Integration architecture (Dong & Srivastava, Big Data Integration, 2015).

2.2 Challenges Involved in Record Linkage of Big Data

Traditional Record Linkage depends on measuring the similarity between a static set of structured tuples that have the same schema. Record Linkage is not a simple and trivial task due to many reasons. First, typographical errors, which prevents records from being associated with the same individual. Second, choosing the fields that will be used in the detecting similarity process between records. Thirds, determining the threshold that will decide which records are considered to be duplicate and which are not (i.e. If the similarity between records is 80% or above they will consider duplicate). In Big Data Integration the situation is more complicated and challenging because data has very heterogeneous structure and coming from many sources (structured, unstructured, and semi-structured), and data sources are continuously evolving and dynamic. This makes the Record Linkage more challenging and non-trivial in Big Data Integration (Dong & Srivastava, Big data integration, 2013).

TABLE I. SUMMARY OF THE THREE BIG DATA INTEGRATION TECHNIQUES (DONG & SRIVASTAVA, BIG DATA INTEGRATION, 2013)

Big Data Property	Schema mapping	Record linkage	Data fusion
Volume	Integrating Deep Web, Web tables/lists	Adaptive blocking, MapReduce-based linkage	Online fusion
Velocity		Incremental linkage	Fusion in a dynamic world
Variety	Dataspace systems	Linking text to structured data	Combining fusion with linkage
Veracity		Value-variety tolerant linkage	Truth discovery

3. RECORD LINKAGE TECHNIQUES THAT HANDLE VOLUME CHALLENGE OF BIG DATA

Traditional Record Linkage approaches become inefficient and ineffective when we are examining large datasets. New techniques have been proposed to address the challenges of volume dimension by using MapReduce to efficiently and effectively parallelize the process of Record Linkage as illustrated in table 1 and explained next. Parallelize Record Linkage depends on effectively distribute the workload between many nodes exploiting the techniques of adaptive blocking (Dong & Srivastava, Big Data Integration, 2015).

Applying Record Linkage with each update of large datasets from scratch is unaffordable especially when the datasets are dynamic and continuously evolving. Incremental Record Linkage has been proposed to address the challenges of velocity aspects. It allows efficient incremental Record Linkage in cases of any updates inserts or deletes of any record in the dataset (Dong & Srivastava, Big Data Integration, 2015).

To address the variety challenge of Big Data Integration, new techniques have been proposed to link or match a structured text with unstructured or free text. Matching structured data with unstructured text could happen in many cases as trying to match unstructured offers to structured products description of people with tweets or blog posts about their shopping experience (Kannan, Kannan, Agrawal, & Fuxman, 2011). Finally, to address the veracity aspects, a variety of clustering and linking techniques have been proposed.

Those techniques focus on out of date values and how to deal with erroneous values effectively. Dealing with erroneous and out of date values is very important because it may prevent correct Record Linkage (Dong & Srivastava, Big Data Integration, 2015). The techniques that handle the increasing volume of data depend on parallel processing where the input data is partitioned into a number of blocks to distribute the workload between them. Then, these blocks will be processed by many numbers of reduce tasks in the MapReduce programming model described below. Map Reduce is not the only way to process Big Data, but Apache Spark has been used to process the data in parallel after distributing it to a number of chunks between the working nodes. Spark is also known as in memory processing techniques, which gains more than 100x faster than MapReduce.

Next sub-sections address Record Linkage techniques that handle the Volume property of Big Data.

3.1 Blocking Techniques

This paragraph discusses the blocking techniques used to partition the Big Dataset into a number of chunks where the expected similar records are more likely to be placed in the same chunk. After distributing the dataset into chunks, a similarity function is executing in parallel between those chunks utilizing the MapReduce programming model. The main idea of parallelism is to speed up the Record Linkage runtime and the exploit the ability to scale up.

The full detection of Record Linkage process requires the complete scan of all records to compute the similarity between them. Performing a complete scan of all records require executing a Cartesian product of similarity checks between n input dataset which will require a complexity of $O(n^2)$. This makes the process of Record Linkage very exhaustive, expensive and tidy process and prone to errors, especially with large datasets. Partitioning the data into blocks where each block contains only records that are more likely to be similar solves this problem by only comparing the records within the same block (Cao, et al., 2011).

Blocking Key is used to partition the Big Datasets into smaller blocks depending on a chosen entity attributes' values. The blocking key is often created by the concatenation of the prefixes of the chosen attributes (Kolb, Thor, & Rahm, Parallel Sorted Neighborhood Blocking with MapReduce, 2011). Standard blocking, Sorted neighborhood, Q-gram Indexing, and Canopy Clustering with TF-IDF are the different Blocking methods for Record Linkage listed by (Baxter, Christen, & Churches, 2003), (Christen, 2011) as illustrated in Table II. The sorted neighborhood method is one of the best-used blocking methods implied by many authors in (Papadakis, Ioannou, Palpanas, Niederee, & Wolfgang, 2013), (Papadakis, Ioannou, Niedere, Palpanas, & NejdI Wolfgang, 2011), (Kolb, Thor, & Rahm, Parallel Sorted Neighborhood Blocking with MapReduce, 2011), . (Kolb, ThorAndreas, & Rahm, Multi-pass Sorted Neighborhood Blocking with MapReduce, 2012), (Mestre & Pires, 2013), (Kolb, Köpcke, Thor, & Rahm, 2011).

3.2 MapReduce Programming Model

MapReduce is a shared nothing-programming model. It is specially designed to handle the exhaustive processing by paralleling distribution of the workload between multiple clusters or multiple nodes. The design of nodes or clusters could be easily scaled out if needed (Dean & Ghemawat, 2004). In MapReduce paradigm, data is partitioned as described before and placed in blocks then, data in each block is represented by key-value pairs. The processing is done to where the data is located using two user-defined functions, Map and Reduce. The map function is used to partition the input dataset into chunks to be processed separately according to the required function. The reduce function is used to sort and collect the final output from each map function and generate one final output.

TABLE II. SUMMARY OF THE BLOCKING TECHNIQUES

Standard Blocking	Sorted Neighborhood	Q-Gram Indexing	Canopy Clustering with TF-IDF
- All records with the same blocking key value (BKV) will be on the same block. - Only records within the same block will be compared to check the duplication between them.	- The database is sorted according to (BKV). - A window of a fixed sized is moved on the sorted records. -Candidate records are generated only from records of the current window.	- All database records that have not only the same BKV but also a similar BKV are inserted into the same block. - Each (Blocking Key Value) is transformed to a list of Q-Grams. - Combinations of these q-gram lists are then generated down to a minimum length, determined by a user threshold. - Then a record identifier is inserted into more than one block.	The clustering is done by calculating Similarity between the Blocking Key Values using measures such as TF-IDF or Jaccard.

Fig. 3 illustrates a simple example of MapReduce program. Suppose we want to calculate the frequency of each char from the input string (3 lines in this case for simplicity) by utilizing MapReduce Programming model. First, the input data is split or partitioned into three chunks. Second, each chunk goes to a mapper, which is responsible to generate a key-value pair (<char>, <char frequency>). Third, the reducer task starts with shuffle and sort step. It sorts the keys generated from the map phase into a larger data list to be easily grouped and integrated into the reduce phase.

By applying the MapReduce model in the Record Linkage problem for Big Data, The Map function will be used to read the input datasets in parallel from the connected nodes and partition them into small chunks and redistribute them based on their blocking key to the reduce functions/tasks that are responsible for Record Linkage process.

3.3 Blocking based Record Linkage using MapReduce programming model

By utilizing MapReduce model to solve the Record Linkage problem with Big Data, The Map function or mapper will read the input datasets in parallel and partition them into small chunks in the form of a key-value pair (blocking key, entity). The default hash partition function will use this key-value pairs to redistribute the blocks based on their blocking key to the reduce functions/tasks. The reducer function is Responsible for Record Linkage process using one or combination of the similarity techniques. Entities will whereas the entity has the same blocking key will be processed in the same reducer.

This guarantees that all input entities having the same blocking key and reside in the same block will be processed using the same reducer as illustrated in Pseudocode depicted in Fig. 4. So, MapReduce offers a convenient programming model for scaling out record Linkage process by running Record Linkage techniques in parallelized blocking using MapReduce model (Hsueh, Lin, & Chiu, 2014). The use of basic MapReduce programming model has the following Limitations:

- It is susceptible to data Skew because of unequal distributions of blocks.

- Single or few number of reducers tasks may be the dominant factor of the execution time of the whole Record Linkage process.
- Some nodes of MapReduce model may be ideal waiting for others to complete their tasks.

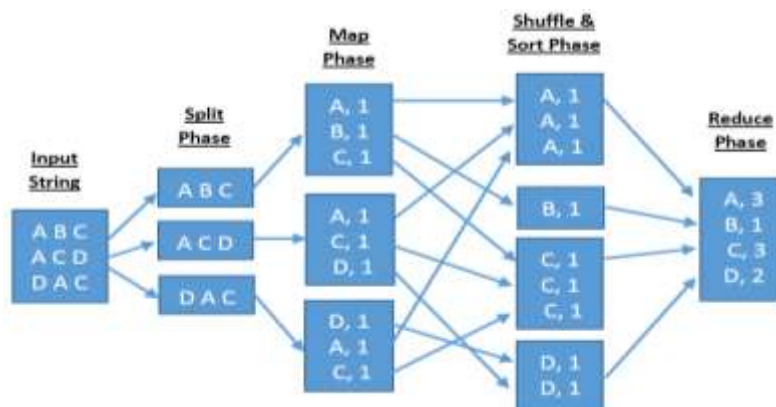


Fig. 3. Calculating the frequency of characters using MapReduce paradigm.

Therefore, the basic MapReduce model with Big Data is not scalable if we don't solve the bottleneck that is happening during the execution of large blocks, not effective because of the time-consuming cost of running large blocks, and finally has difficulty of handling fault tolerance.

Partitioning Big Datasets into smaller blocks using blocking key techniques may lead to unequal distributed blocks which cause a problem known as data skew. Data skew is happening when the data is not evenly distributed across data partition. An uneven distribution of data degrades the performance of the overall execution as some nodes will sitting idle waiting for some other to finish their job with larger sizes. Data skew can occur on both map and reduce phases. Solving data skew that happens in map phase is easy whereas it is very complex and demanding issue in the reduce phase because the majority of the execution is happening in the reduce phase.

A small number of studies consider the use of MapReduce paradigm to solve the load-balancing problem, especially in the Record Linkage. Hadoop is one of the most famous frameworks that implement MapReduce programming model. An extremely improvement in the efficiency of the whole process of the Record Linkage is noticed by balancing the workload of the reducers in the matching step (Hsueh, Lin, & Chiu, 2014).

Unequal distributed of the blocks will cause many problems as Load balancing and fault tolerance between the Reducers in MapReduce programming Model. Many authors tried to solve the load balancing problem by utilizing the idea of blocking key distribution as Yipeng (Huang, 2011) introduced a partitioning solution that dynamically balances the workload of records' comparison to solve the problem of uneven workload that happens due to unequal partitions blocks. Their evaluation achieves a significant improvement that outperforms the default partition of Hadoop for Record Linkage problems involving data skew.

Kolb, et al. (Kolb, Thor, & Rahm, Block-based Load Balancing for Entity Resolution with MapReduce, 2011) addressed the skew problems by proposing a general balancing

approach named BlockSplit. The aim of BlockSplit is to reduce the search space of Record Linkage and evenly distribute the workload between the reducers in MapReduce programming model. BlockSplit consists of 2 main jobs. Each of them is done by a number of maps and reduces tasks. First one is analysis job. This job receives the input partitions and performs blocking using blocking key. The output of this job is the number of entities per each block, known as Block Distribution Matrix (BDM). BDM is the input of the second job known as matching phase. Record Linkage is done between elements of the same block in this phase. The approach of BlockSplit took the size of blocks into account and assigned each block to a number of reduce tasks after passing the checking of load balancing constraints. Small blocks will be processed in one reducer. Large Blocks are divided into smaller numbers of sub-blocks based on the input partitions for better performance of parallel matching occurred within multiple reduce tasks. But this approach as mentioned by Kolb, et al. is (Kolb, Thor, & Rahm, Load Balancing for MapReduce-based Entity Resolution, 2012) may still lead to unbalanced reduce load balance because of the different size of sub-blocks. Also, they used round robin as a balancing technique in the first job (analysis) in order to evenly distribute the blocks between the available reducers. Using round-robin technique did not lead to a balanced number of elements between reducers, although it could be ignored because time in this job is not the dominant factor.

```

MapO
{
    Foreach input entity
    {
        Determine the blocking key
    }
    Return Key-value pairs (blocking key, entity)
}
Use the default hash partition to assign each block to the ReduceO
Foreach Block
{
    Reduce O
    {
        Calculate Similarity between entitiesO
        Return matching entity pairs
    }
}
    
```

Fig. 4. Pseudocode of Record Linkage of Big Data using basic MapReduce model.

Kolb, et al. (Kolb, Thor, & Rahm, Load Balancing for MapReduce-based Entity Resolution, 2012) proposed and evaluated PairRange, a load balancing for MapReduce based Entity resolution approach. PairRange redistributes the entities between reducers such that each reducer computes nearly the same numbers of entities comparisons. PairRange consists of two main jobs, first one is calculating the BDM as described in (Kolb, Thor, & Rahm, Block-based Load Balancing for Entity Resolution with MapReduce, 2011). The second job calculates the total number of comparisons from all blocks and redistributes them to all available reducers such that each reducer takes almost an equal range of entities comparisons. The approaches proposed by Kolb, et al. in (Kolb, Thor, & Rahm, Block-based Load Balancing for Entity Resolution with MapReduce, 2011) and (Kolb, Thor, & Rahm, Load Balancing for MapReduce-based Entity Resolution, 2012) are scalable as they scale for increasing the number of working nodes comparing to the basic MapReduce model, Robustness against data skew and finally achieve a decrease in execution time comparing to Basic MapReduce model as many nodes are added. BlockSplit and PairRange approaches are only applied to a large scale of

datasets and not yet explore in Big Datasets. No filtering or pruning strategy where implemented in both approaches to limit the number of comparison between elements of the same block. In pairRange approach, Elements are not equally distributed in the reducers although the resulting comparisons are balanced between reducers. It is not clear in PairRange approach how it ignores many elements comparisons as it occurs in other reducers.

Chen, et al. (Chen, Pullen, Petty, & Talburt, 2015) introduced a methodology to carry out Record Linkage without having to perform pair-wise matching. They used matching keys based on MapReduce framework. This methodology consisted of three main parts, Standardization, match key generation, and transitive closure. They overcome the limitation of transitive closure by designing a new iterative transitive closure that applies the method on multiple match keys in MapReduce scenario.

Papadakis, et al. (Papadakis, Ioannou, Niederée, Palpanas, & Nejd, 2011) proposed a method to remove redundant comparisons. They also introduced a measurement for quantifying the redundancy entailed by blocking method and explain how to use it to tune the process of comparison pruning. They applied the proposed blocking techniques on two large datasets. The results showed a remarkable increase in the efficiency of the comparison and blocking process.

While Jin, et al. (Jin, et al., 2013) proposed DISC, a distributed algorithm used for single Record Linkage hierarchical clustering based on MapReduce programming model. They introduced the analysis of the algorithm, including an upper bound on the computation cost. The Algorithm was flexible to run in a big dataset by configuring some parameters and showed a scalable speedup of up to 160 on 190 computer cores.

Hsueh, et al. (Hsueh, Lin, & Chiu, 2014) applied multiple key distributions. They tried to solve the Record Linkage problem for a large dataset by proposing a MapReduce algorithm that has multiple keys. The proposed algorithm consists of two phases, first is the combination-based blocking and second is load-balanced matching. They used the two keys in the combination-based blocking to filter out unnecessary entity pairs. In the matching step, the load balancing is done by obtaining statistics of the total number of computation required for each block than evenly distribute the number of comparisons between all the reducers. They only used one input partition. Solving the load balancing problem in this case is much easier than having more than one input partition as in (Kolb, Thor, & Rahm, Block-based Load Balancing for Entity Resolution with MapReduce, 2011) and (Kolb, Thor, & Rahm, Load Balancing for MapReduce-based Entity Resolution, 2012).

Current state of art employs the supervised or unsupervised learning based approaches in the Record Linkage problem where supervised learning aims to classify a pair of records as matched or non-matched. The classifier is learned or trained depending on sets of records labeled as matched or non-matched then it is tested until having an acceptable accuracy. The goal of learning based Record Linkage is to classify new pair records as matched or non-matched depending on the trained data.

Many authors considered the use of machine learning in the Record Linkage problem as Kolb, et al. (Kolb, Köpcke, Thor, & Rahm, 2011) discussed the use of MapReduce programming model to parallel the workload distribution between many reducers using

variations of the Sorted Neighborhood Method (SNM) with a varying size window. They proposed two different strategies for calculating similarity in the case of the Cartesian product of two input sources. The first one named MapSide depended only on computing similarity in the Map phase. While the second one name ReduceSplit evenly distributes entity pairs of the two input sources across all available reduce tasks that finally apply similarity computation for each reducer. Each pair-wise similarity between the common properties in the two data sources serves as a feature of the classification model. Computing similarities are the dominant factor in the classification model because it consumes about 95% of the time. No blocking techniques or load balancing strategies were employed to handle the computational skew results because of unbalanced blocks. No pruning similarity computation was developed to reduce the overall number of comparisons of entities that are likely to be non-matched.

Kolb, et al. (Kolb, Thor, & Rahm, Dedoop: Efficient Deduplication with Hadoop, 2012) devolved a tool called Dedoop (Deduplicate with Hadoop) which supports MapReduce-based Entity Resolution for large datasets. Dedoop had a GUI for its workflow as shown in Fig. 5 that consisted of three steps; blocking, similarity computation and matching decision of the input blocks depending on their similarity value. The final step (match classification) could depend on a trained classifier using a Machine Learning algorithm on the training dataset. Several blocking techniques could be used in the blocking, similarity computation, and match classification. Dedoop did not offer any pruning steps to reduce the number of unnecessary comparisons.

Cao, et al. (Cao, et al., 2011) proposed a new algorithm for the blocking of the Record Linkage problem. The proposed algorithm learned the blocking schema for both labeled data and unlabeled data. Experiments showed that using unlabeled data in the learning process could decrease the number of candidate matches while at the same time maintaining the same level of true matches (Cao, et al., 2011).

Moir, et al (Moir & Dean, 2015) introduced the Generic Record Linkage (GRL) framework to classify pairs of entities as matching pairs or non-matching pairs based on their features and the semantic relationship between them. The proposed approach applied supervised Machine Learning to determine the features of the attributes and their semantic relationships used in the classification process. The applying of the proposed approach results in increasing the accuracy of the classification.

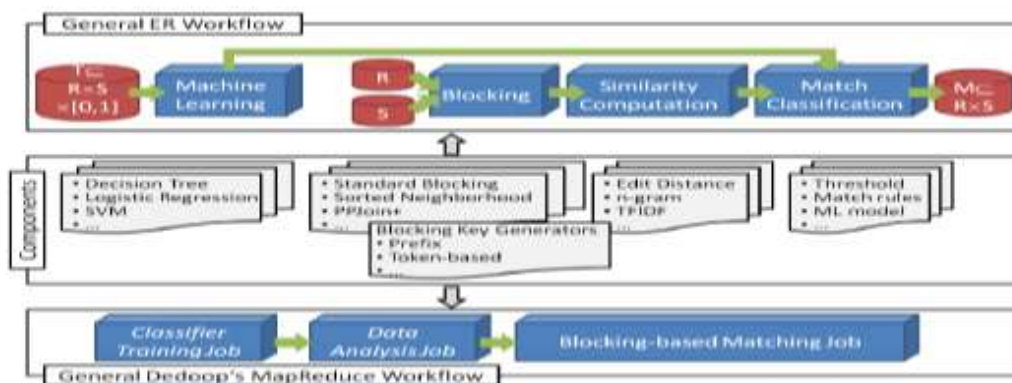


Fig. 5. Overview of Dedoop (Kolb, Thor, & Rahm, Dedoop: Efficient Deduplication with Hadoop, 2012).

3.4 Spark:

Apache Spark is a fast clustering computing technology built on top of MapReduce framework. It extends the functionality of MapReduce model to efficiently handle big data stream, as it is an in-memory clustering computation. Handling data streams in memory helps to increase the processing speed of the applications, as there is no time spent on moving the data to and from disk. Spark provides an API for many programs as Java, Scala, and Python. Spark provides a layer for a large number of tools such as Mlib for machine learning, Structure data processing, Spark SQL, GraphX for graph processing and Spark streaming for stream processing (Shoro & Soomro, 2015).

Spark is a framework that based on the design of Resilient Distributed Dataset (RDD). There are two types of RDD operations: transformations and actions. Transformation is a function, which produces a new RDD from the existing one. There are many functions of transformations such as map(), unions(), distinct(), groupby(), key(), sortBykey() and join(). Actions are following transformations and return results to the driver program. There are many functions of actions as collect(), Count(), reduce(), take(), CountByValue(), foreach(), and aggregate(). Transformations are lazy executions that mean they should be followed by actions to take place. In-memory computation is the main advantage of Spark, where it means that intermediate data is kept in memory, which will allow for 100 times faster than Hadoop MapReduce paradigm (Pita, et al., 2015).

Pita et al. (Pita, et al., 2015) presented Spark based framework of probabilistic record linkage. The proposed framework integrated three databases from different sources. They also used a bloom filter technique for anonymization before applying linkage stage. The framework consisted of four phases:

- Phase 1: data quality assessment. Aims to identify the attributes more suitable for probabilistic record linkage. The attributes were chosen based on their co-existence in the DB, their discriminatory capabilities, and their quality in terms of missing values or null. They used SPSS for achieving these tasks.
- Phase 2: Pre-processing. Aims to apply data transformation and standardization to prepare the dataset for record linkage process such as transforming to uppercase, substitute null values with pre-defined values. For data privacy, bloom filter has been applied to anonymize the fields before the record linkage phase takes place. Spark is used for performing the transformation and standardization tasks.
- Phase 3: Linkage process. A deterministic record linkage had been applied between two DB using direct join and probabilistic record linkage where Spark was used to perform record linkage.
- Phase 4: Evaluate the data mart. SPSS was used to evaluate the produced data marts.

4. CONCLUSION AND FUTURE WORK

Record Linkage techniques have many challenges involved in Big Data due to the high resolution of it. This paper surveys the Record Linkage techniques used to solve the volume property of Big Data. Techniques used MapReduce model and Spark models for distributed computing and parallel processing. All solutions depend on partitioning the data into blocks using a chosen key by map function, and then resolve the duplicates using one of the techniques of duplicate resolution in reduce function. Spark is more efficient due its in memory processing capabilities, which make it faster more than 100 times than

MapReduce. Partitioning input data into blocks may lead to data skew, which results in unbalanced workload that can be solved by additional MapReduce Job to determine the blocking key distribution.

Yet there is not much effort done in load balancing of MapReduce techniques so more work should be dedicated to it because it has not been actively explored. More attention should be pointed to pruning an unnecessary number of comparisons of entities that are likely to be non-matched during the similarity computations. In addition, using ML in the field of Record Linkage is still in its infant stage and needs more efforts to be exerted on it. In future work, we plan to find an adaptive method to balance the load between the data blocks in MapReduce programming model. We will try to execute the mentioned load balancing approaches (BlockSplit and PairRange) on Big Data as they only explored on large data scale. We will exploit ML to formalize the best attributes that could be used for blocking. In addition, trying to apply some pruning methods to avoid the unnecessary comparisons between the records located in the same blocks.

REFERENCES

- Baxter, R., Christen, P., & Churches, T. (2003). comparison of fast blocking methods for record linkage. *ACM SIGKDD '03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*.
- Cao, Y., Chen, Z., Zhu, J., Yue, P., Lin, C.-Y., & Yu, Y. (2011). Leveraging Unlabeled Data to Scale Blocking for Record Linkage. *Proceedings of the 22nd International JointConference on Artificial Intelligence (IJCAI)*.
- Castanedo, F. (2015). *Data Preparation in the Big Data Era*. USA: O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA.
- Chen, C., Pullen, D., Petty, R. H., & Talburt, J. R. (2015). Methodology for Large-Scale Entity Resolution Without Pairwise Matching. *IEEE International Conference on Data Mining Workshop (ICDMW)*.
- Christen, P. (2011). A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. *IEEE Transactions on Knowledge and Data Engineering X(Y)*.
- Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. *the 6th conference on Symposium on Operating Systems Design & Implementation*. Berkeley, CA, USA.
- Dhavapriya, M., & Yasodha, N. (2016). Big Data Analytics Challenges and Solutions Using Hadoop, Map Reduce and Big Table. *International Journal of Computer Science Trends and Technology (IJCTST)*, 4(1).
- Dong, X. L., & Srivastava, D. (2013). Big data integration. *ICDE*, (p. 1245:1248).
- Dong, X. L., & Srivastava, D. (2015). *Big Data Integration*. Morgan & Claypool Publishers.
- Getoor, L., & Machanavajjhala, A. (2012). Entity Resolution: Theory, Practice & Open Challenges. *VLDB Endowment* 5(12) (pp. 2018-2019). ACM.
- Hsueh, S.-C., Lin, M.-Y., & Chiu, Y.-C. (2014). A Load-Balanced MapReduce Algorithm for Blocking-based Entity-resolution with Multiple Keys. *Parallel and Distributed Computing*, (p. 3).
- Huang, Y. (2011). *Record linkage in an Hadoop environment*. School of Computing, National University of Singapore.
- Jin, C., Patwary, M. M., Agrawal, A., Hendrix, W., Liao, W.-k., & Choudhary, A. (2013). DiSC: A Distributed Single-Linkage Hierarchical Clustering Algorithm using MapReduce. *4th International SC Workshop on Data Intensive Computing in the Clouds (DataCloud)*.
- Kannan, A., Kannan, A., Agrawal, R., & Fuxman, A. (2011). Matching unstructured product offers to structured product specifications. *17th ACM SIGKDD international conference on Knowledge discovery and data mining*. (p. 404:412). ACM.
- Kolb, L., Köpcke, H., Thor, A., & Rahm, E. (2011). Learning-based Entity Resolution with MapReduce. *CloudDB*, (p. 1:6).
- Kolb, L., Thor, A., & Rahm, E. (2011). Block-based Load Balancing for Entity Resolution with MapReduce. *Proceedings of the 20th ACM international conference on Information and knowledge management*, (p. 2397:2400).

- Kolb, L., Thor, A., & Rahm, E. (2011). Parallel Sorted Neighborhood Blocking with MapReduce. *Proc. Conf. Datenbanksysteme in Büro, Technik und Wissenschaft* (pp. 45–64). BTW.
- Kolb, L., Thor, A., & Rahm, E. (2012). Dedoop: Efficient Deduplication with Hadoop. *VLDB Endow*, 12(5), 1878–188.
- Kolb, L., Thor, A., & Rahm, E. (2012). Load Balancing for MapReduce-based Entity Resolution. *International Conference on Data Engineering (ICDE), IEEE*, (pp. 18-629). Leipzig, German.
- Kolb, L., Thor, A., & Rahm, E. (2012). Multi-pass Sorted Neighborhood Blocking with MapReduce. *Computer Science-Research and Development*, 27(1), pp. 45–63.
- Kong, C., Gao, M., Xu, C., Qian, W., & Zhou, A. (2016). Entity Matching Across Multiple Heterogeneous Data Sources. *International Conference on Database Systems for Advanced Applications*. Cham: Springer.
- Köpcke, H., Thor, A., & Rahm, E. (2010). Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, 3, pp. 484–493.
- Mestre, D. G., & Pires, C. E. (2013). An Adaptive Blocking Approach for Entity Matching with MapReduce. *SBB*.
- Moir, C., & Dean, J. (2015). A Machine Learning approach to Generic Entity Resolution in support of Cyber Situation Awareness. *Proceedings of the 38th Australasian Computer Science Conference (ACSC 2015)*, 27, p. 30.
- Papadakis, G., Ioannou, E., Niedere, C., Palpanas, T., & NejdlWolfgang. (2011). Eliminating the Redundancy in Blocking-based Entity Resolution Methods. *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries* (pp. 85-94). ACM.
- Papadakis, G., Ioannou, E., Niederée, C., Palpanas, T., & Nejdl, W. (2011). To Compare or Not to Compare: Making Entity Resolution more Efficient. *Proceedings of the international workshop on semantic web information management* (p. 3). ACM.
- Papadakis, G., Ioannou, E., Palpanas, T., Niederee, C., & Wolfgang, N. (2013). A Blocking Framework for Entity Resolution in Highly Heterogeneous Information Spaces. *IEEE Transactions on Knowledge and Data Engineering*, 25(12), pp. 2665–2682.
- Pita, R., Pinto, C., Melo, P., Silva, M., Barreto, M., & Rasella, D. (2015). A Spark-based workflow for probabilistic record linkage of healthcare data. *EDBT/ICDT Workshops*, (pp. 17–26).
- Shoro, A., & Soomro, T. R. (2015). Big Data Analysis: Apache Spark Perspective. 15. *Global Journal of Computer Science and Technology: C Software & Data Engineering*.

Overview of Integrating Granular Computing with Big Data

Osama Sayed Abdel-Rahman *

Hesham Ahmed Hefny*

Abstract.

Recently, "Big Data" is a buzzword which encourages data scientists, data analysts, and researchers to extend their work and modify their techniques to overcome the new challenges come with big data concepts. The aim of this paper is to introduce the challenges and issues comes with big data in the domain of data mining and present the new techniques to overcome these Issues and challenges. As one of these new techniques, we proposed in this paper an overview of adopting the integration between granular computing (in particularly fuzzy set theory, rough set theory, and data mining) with big data to develop a new technique for mining big data and take advantages of both research areas to extract useful rules from huge data to enhance the decision making process.

Key words : Granular Computing - Rough sets – Fuzzy sets – Big Data

1. Introduction

Nowadays, with the very large amount of generated data from different sources and devices, the process of extracting valuable information and nontrivial knowledge of this huge amount of data becomes one of the important and interesting research areas. Recently, big data analysis and mining is attracting more attention for both scientific and industrial areas.

In the last decade, a new high-level programming model called MapReduce is firstly introduced by Google and widely adopted for parallel and distributed computing. Apache Hadoop is the first open source implementation of the MapReduce programming model introduced by Yahoo and become more popular because Hadoop allows developers very easily to create many functions such as *map()* , *reduce()* . The first Hadoop Function *map ()* processes massive data set partition in parallel while *reduce ()* aggregates the results. The Hadoop Distributed File System (HDFS) is a disk-based file system that spans across multi nodes of a distributed system. Files stored in HDFS are automatically divided into blocks, replicated and distributed to the different nodes of the local disks [Koliopoulos et. al , 2015].

* Department of Computer and Information Sciences - Institute of Statistical Studies and Researches - Cairo University

Hadoop is not optimal for big data analysis algorithms which need iterative processing and repeated reading for the same data set, due this limitation Apache Spark was firstly introduced as a new open source implementation of MapReduce programming model. Spark keeps the benefits of Hadoop such as parallel processing, load balancing, and fault tolerance in addition to introducing a new benefit which is storing the intermediate results into main memory. This new feature makes spark 10-100 times faster than Hadoop. Due these capabilities of Spark, Apache Spark recently gains an increasing attention from developers and researchers which make it the favorite big data parallel computing framework [Alkowaileet et. al , 2015], [Park and Peng , 2016], [Naacke et. al , 2016].

Although Apache Spark is considered as powerful large scale data processing analytics engine but it doesn't provide advanced indexing storage engine because Apache Spark depends on HDFS (Hadoop Distributed File System) which is considered as a limitation of Spark. To overcome this limitation, Spark can integrate with other database engines which introduce indexing capabilities such as MongoDB, Apache HBase, and Apache Cassandra [Alkowaileet et. al , 2015].

Granular computing has emerged as a new rapidly growing information processing paradigm inside the community of Computational Intelligence. The basic idea of granular computing is that any subset of the universe or any equivalent class of the universe, or and module of a system can be viewed as granules and the granules can be decomposed into smaller granules called sub-granules [Yao et. al , 2013].

The main objective of this paper is to demonstrate the power of integrating big data technologies and techniques with granular computing. In particularly, integrating fuzzy logic concepts with rough sets theory methodology of data mining with big data to enhance big data mining techniques. Such an approach helps in extracting hidden knowledge in big data. This objective will be achieved in two steps, first is to use fuzzy logic for fuzzifying numerical values big data set, and second is the integration with relational database systems.

The rest of this paper is organized as follows. Section 2 introduces the basic background of Granular Computing and big data technologies. In section 3 we discuss big data mining challenges and issues. In section 4, we proposed the basic idea of integration between granular computing and data mining. the paper ends with conclusion in section 5.

2. Preliminaries

2.1 Granular Computing

Granular computing is a rapidly growing research area which provides a great conceptual framework for data mining. Granular computing has emerged as a new rapidly growing information processing paradigm inside the community of Computational Intelligence. The buzzword "Granular Computing" is firstly introduced in 1997, but the main ideas of granular computing have been studied in many research fields with different names. Rough sets theory, fuzzy sets theory, data analysis, data mining, and machine learning are the most famous names of these fields which use granular computing principals [Yao , 2006].

Granular computing is considered as a new field of studies deals with theories, methodologies and techniques which use granules in the process of problem solving. Granular computing can be defined as general problem solving theory based on various levels of granularity and details. The first studies of granular computing started inside the computational intelligence community. Zadeh was firstly introduces the notation of information granulation in 1979 and he suggested that his theory of fuzzy sets can find potential application in this respect. In 1982 Pawlak has introduces rough sets theory which provide solid example of granular computing [Yao , 2006].

After a decade, In 1997 Zadeh recall information granulation to renew the researcher's interest [Zadeh, 1997]. Lin was firstly suggesting the buzzword "Granular Computing" for this green research field in 1997 [Lin, 1997].

In 2010 Yao has introduced granular computing as a human-inspired problem solving paradigm. He has explored the multi-level, hierarchical structures which makes granular computing a human-inspired and structured approach to problem solving process [Yao , 2010]. Figure 1 illustrates multi-view and multilevel granular structure.

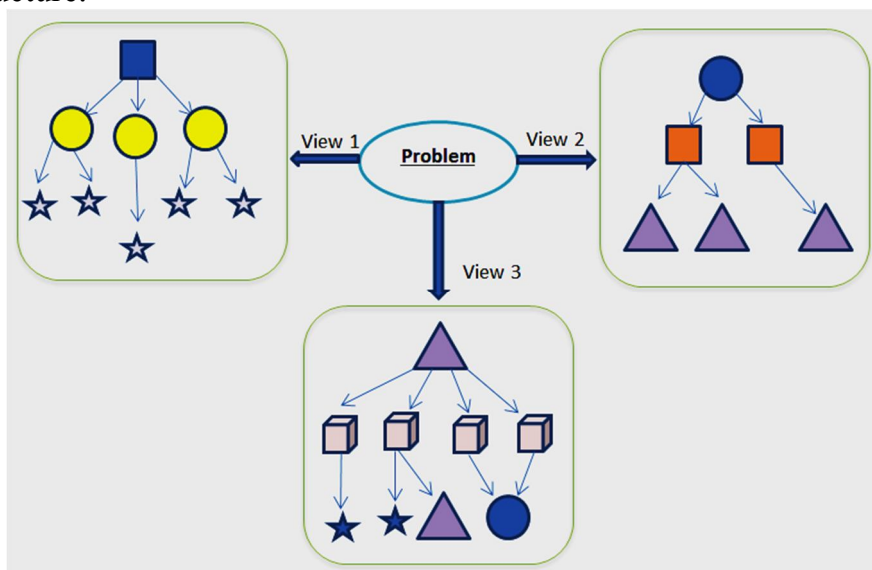


Figure 1 Multi-view and multi-level granular structures

In [Yao , 2010], [Yao , 2008] Yao describes Triarchic theory of granular computing as a unified view consists of three integrated levels of structures as shown in Figure 2.

- 1- Philosophical level focus on structured thinking
- 2- Methodological level focus on structured problem solving
- 3- Computational level focus on structured information processing

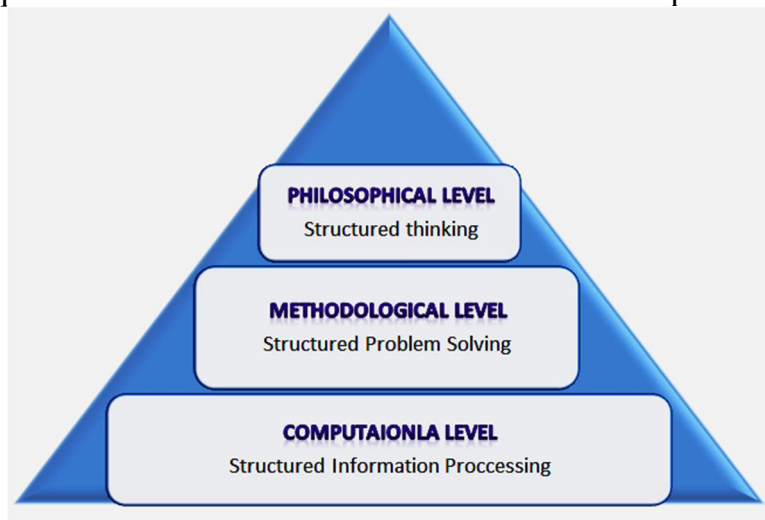


Figure 2 the granular computing triangle

As a base for data mining, granular computing can help for extracting rules. Rules are one of the most useful representations of knowledge discovery from huge data. One of the most important key notations of fuzzy sets theory is linguistic variables which is considered as fuzzy granule and can be represented by natural language. By using fuzzy granule "linguistic variables" rules become more human friendly and more understandable and useful because it provide a smooth transition between member and non-member of a set which is easily understandable to human.

Rough sets theory also applied to data mining as a strong example of granular computing. In order to mine more meaningful rules we can group attribute values into granules. In the other hand granular computing can be combined with other data mining methods to get more effective and efficient methods for extracting more useful rules from huge amount of data sets.

In 2016 Skowron et al. introduces interactive granular computing to help decision support system researchers in solving complex problems by constructing granules. To do so Skowron extends the existing granular computing approach to interactive granular computing by introducing complex granules [Skowron et. al , 2004].

2.2 Big Data Tools and Technologies

2.2.1 Hadoop Ecosystem

Google in 2004 firstly introduced the MapReduce model which is a high-level programming model and implementation for distributed computing model deal with large scale parallel data processing by Jeffrey Dean and Sanjay Ghemawat [Dean and Ghemawat, 2004]. MapReduce model uses two main functions map and reduce. Map processes massive data set partition in parallel while reduce aggregates the results. MapReduce model consists of one master node and multi worker nodes. The master node is responsible for handling data partitioning and tasks scheduling between workers nodes.

In 2005 Yahoo introduced Apache Hadoop as an open source MapReduce implementation. The Hadoop Distributed File System (HDFS) is a disk-based file system that spans across multi nodes of a distributed system. Files stored in HDFS are automatically divided into blocks, replicated and distributed to the different nodes of the local disks.

HDFS maintains metadata about the location of blocks and helps Hadoop to schedule each node to process local blocks as well as receive blocks through the network. HDFS encapsulates distributed local storage into a single logical unit and automates the procedure of distributed storage management. Although MapReduce can express many data mining algorithms efficiently, significant performance improvement is possible by introducing a loop-aware scheduler and main memory caching. Data mining algorithms tend to involve multiple iterations over a dataset and thus, multiple, slow, disk accesses. As a consequence, storing and retaining datasets in-memory and scheduling successive iterations to the same nodes should yield significant benefits.

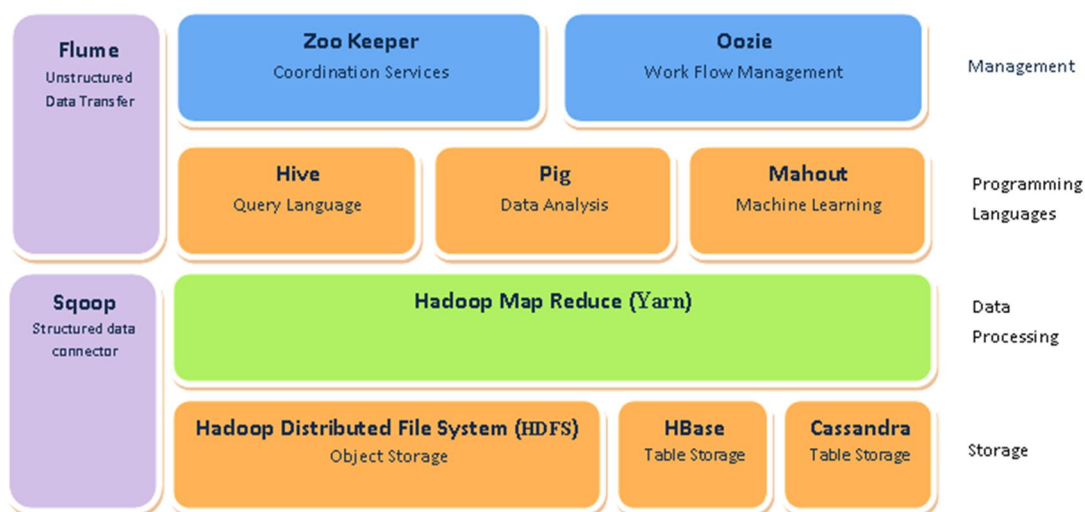


Figure 3 Hadoop Ecosystem

Table 1 Summary of Hadoop Ecosystem Functionality

Hadoop Ecosystem Component	Main Function
HDFS	Object Storage (Unstructured data)
HBase	Table Storage (Structured data)
Cassandra	Table Storage (Structured data)
Hadoop Map Reduce (Yarn)	Distributed Data Processing
Hive	Query Processing
Pig	Data Analysis
Mahout	Machine Learning
Sqoop	Structured Data Connector (Export and Import Data)
Flume	Unstructured Data Transfer (ETL Tool)
Zoo Keeper	Coordination Services
Oozie	Work Flow Management and Scheduling

2.2.2 Spark Ecosystem

Apache Spark has been introduced in 2010 at University of California, Berkeley by Matei Zaharia et al., [Zaharia et. al , 2010]. Spark supports main-memory caching and possesses a loop-aware scheduler. Spark provides in-memory computing capabilities to deliver speed. Additionally, Spark as well as Hadoop implements the MapReduce paradigm and is Java-based. These features enable users to deploy existing Hadoop application logic in Spark via many APIs for ease of development like Scala, Python, and Java to support a wide variety of applications. While MapReduce basically is designed to develop batch processing using disks as intermediate storage, Spark enables near-real time development by using memory for processing and sharing information efficiently. [Katsipoulakis et. al , 2015].

Spark controls a set of libraries including MLlib, Spark SQL, GraphX, Spark Streaming, and Spark R. The basic data structure of Spark is Resilient Distributed Data sets (RDD). RDD is fault tolerance work in parallel figure 4 illustrates Spark ecosystem.

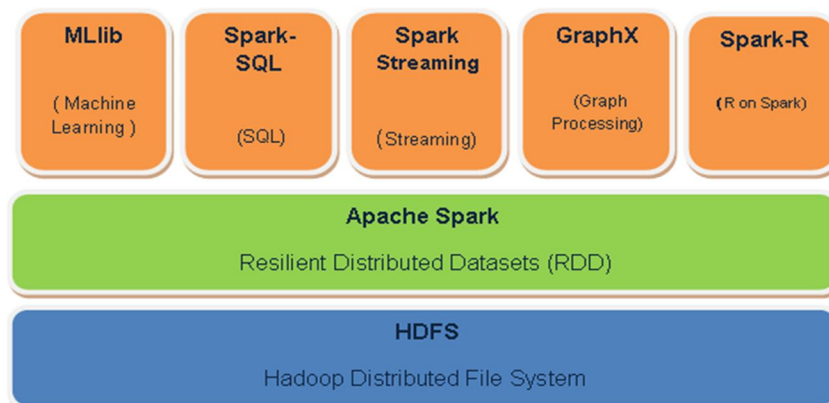


Figure 4 Spark Ecosystem

Spark SQL is a Scala implementation of big data query processing system built on top of Spark. Spark SQL act as SQL Interface and programmatic interface to fill the gap between machine learning and analytics quires and considered as Spark structured data processing module which use a distributed collection of rows called data frame figure 5 illustrates Spark SQL Interfaces.

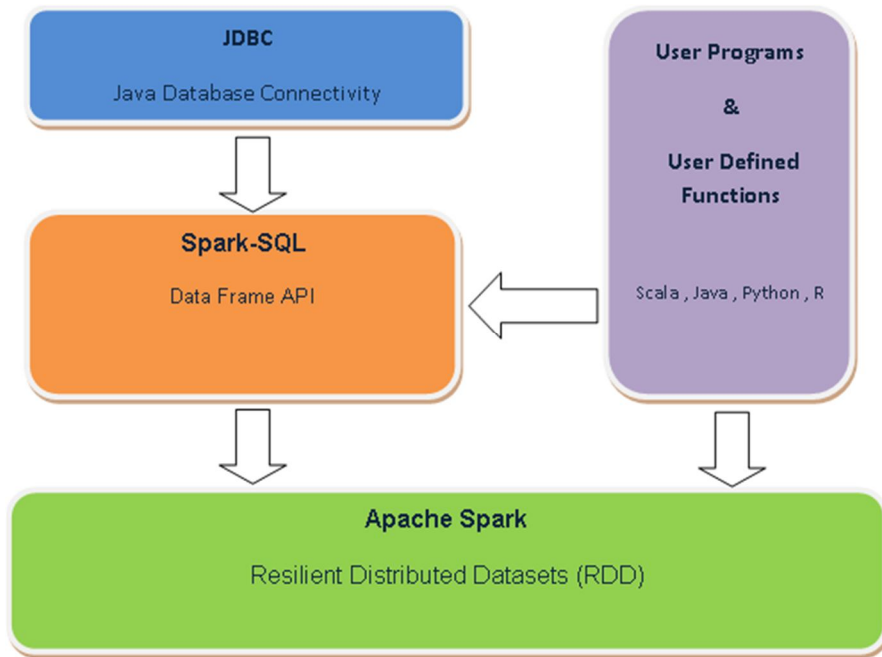


Figure 5 Spark SQL Interfaces

The key point of Spark SQL is data frames which is a distributed collection of rows with the same schema. Data frames are equivalent to table in traditional database systems. Data frames can be represented by using columnar data format which allow to access specific column without need to retrieve the whole data set.

Spark SQL data frames is loaded into main memory instead of reading data from disk , so that the processing is more faster when compare with other big data processing systems , especially for iterative data processing . For all these reasons Spark SQL is considered as high-speed big data query processing system [Park and Peng , 2016]

3. Data Mining Issues and Challenges with Big Data

Opportunities always bring challenges. When we entered big data era we gained many benefits for both business and society. These benefits come with many challenges. One of the most important challenges is the need for adopting new algorithms and approaches for mining and analyzes big data.

Mining big data face lots of issues and new challenges as well as existing data mining techniques face great difficulties when they are required to handle the heterogeneous data sources , large volume data, very speed, privacy, and accuracy coming with big data and big data mining. These challenges in systems and platforms level, Algorithms level, and Business modeling level. The key issues and challenges are variety, volume, Speed, accuracy, privacy [Wani and Jabin, 2018].

3.1- Heterogeneity (Variety)

Classical data mining techniques have been used to discover useful knowledge from structured, homogeneous datasets. Nowadays big data generated from different sources leading to heterogeneity or variety of big data and possesses many types and forms of data; also it is interrelated, interconnected and inconsistently represented. Mining such datasets is great challenge and the degree of complexity is not imaginable before we get there. The heterogeneity of big data means a new opportunity of discovering knowledge or hidden patterns at the intersections within heterogeneous big data [Pal and Saini, 2014], [Kumar and Kumari, 2015].

3.2 - Scalability

The huge volume of increasing generated data requires high scalability data management and high scalability data mining tools. Cloud computing and parallel computing architectures provide great scalability for dealing with the volume challenge of big data. Big data mining implies more time consuming navigation in a massive search space, and prompt feedback from users must be beneficially exploited to help decision making process [Pal and Saini, 2014].

3.3 - Velocity (Speed)

Fast accessing, managing, and mining big data is highly essential in big data mining process. Processing and mining of a task must be finished within a definite period of time; otherwise, the data mining process results becomes less valuable or worthless. Speed is also relevant to data size the larger the datasets the more time for analyzing it. Data mining speed depends on data access time and the efficiency of the mining algorithms. Design of new and more efficient indexing schemes is much desired, but remains one of the greatest challenges to the research community. Controlling the probable parallelism in the access and mining algorithms can improve the speed of big data access and mining [Pal and Saini, 2014].

3.4- Accuracy

With big data, the data sources are of many different origins, not all well-known, and not all confirmable. As a result, the accuracy and trust of the source data quickly become a serious concern, which also affects the mining results.

3.5 - Privacy

Data privacy has become extremely serious concern with mining big data because it often requires personal information to produce relevant results such as location-based and personalized services. Also, with the huge amounts of big data such as social media data which contains great amount of highly interrelated personal information, each bit of information can be mined out. Every transaction regarding our daily life is being pushed to online and leaves a trace there personal communications such as emails, instant message, blogs, and tweets besides online shopping information. As time goes, every piece of your personal information will be scattered here or there. Everyone would easily gain the privilege of using powerful tools to extract your confidential information. Evaluation and prevention of privacy violation during knowledge mining are two related concerns that call for serious investigation and novel solutions [Pal and Saini, 2014].

3.6 - Users Collaboration

Although the great advances in computational analysis there is need for human collaboration. Some complex patterns need to overcome by multiple experts from many domains. Big data mining systems must support input from multiple experts from different locations and shared results [Mall and Rana, 2016].

3.7 – Lake of Big Data Professionals

Big data come with new tools and algorithms such as MapReduce, Hadoop, Spark, Mahout, HBase, and HDFS and so on. All these tools and programming models need professional people to deal with it efficiently and effectively to get best performance [Wani and Jabin, 2018].

3.8 - Garbage Mining

Garbage mining is a serious research topic. In Cyberspace data can quickly become obsolete, contaminated, and inadequate. In addition, there are data created as junks like junk emails. Therefore great attentions and research efforts are required for having a relatively clean cyberspace. Cyberspace cleaning is not an easy task because garbage is hidden, and there is an ownership issue. So in the new era of big data mining for garbage and cleaning it is very important as well as data mining and remains one of the greatest challenges in big data mining [Pal and Saini, 2014].

3.9 - Building efficient big data mining platform

To handle big data and its characteristics, an efficient big data Processing and computing framework is needed. Traditional data mining algorithms only needed all the data to be loaded into the main memory and perform the operation of data mining. In medium scale data processing, parallel computing is used with limited number of processors. As big data applications are characterized by autonomous sources and decentralized controls, consolidating distributed data sources to a centralized node for mining is systematically discouraging due to the potential transmission cost and privacy issues. So building an efficient platform to mine big data is essential and needed [Sherin et. al , 2014].

3.10 - Building efficient big data mining algorithms

With the exponential growth of data, traditional data mining algorithms have been unable to meet large data processing needs. In order to deal with big data, an efficient model that deals with cost effective computation of huge, heterogeneous, sparse, incomplete, complex data are needed. The main drawbacks of big data mining algorithms are lack of user-friendly interaction support, quality and performance. Data mining algorithms usually needs scanning of entire data for obtaining perfect statistics and there may require intensive computing. Therefore it is essential to improve the efficiency and performance of data mining algorithms to handle big data [Sherin et. al , 2014].

3.11 - Visualization

Representing knowledge in good visual format helps the decision maker and improves the decision making process. Although currently there are some tools for visualizing the big data, with the rapidly increasing of data very soon in the near future the existing tools will face difficulties to deal with this huge data. Researchers must pay more attention in visualizing big data considering it as new challenge [Wani and Jabin, 2018].

4. The Proposed Approach

There is urgent need for some methods and approaches which have the capability to overcome the big data mining challenges. one of these important challenges is adopting new algorithms and approaches to deal with big data. We propose one of these approaches by integrating the power of MapReduce model , Hadoop , Spark , and Spark SQL with granular computing especially rough sets theory and fuzzy sets theory in big data mining field. Adopting granular computing in big data mining helps us to focus on high conceptual level and ignore irrelevant and trivial details. Such an approach can help us to extract very useful information hidden in huge data.

5. Conclusion

In this paper, we have presented an overview of granular computing as a powerful view of modeling real world problems especially complex ones. We introduce big data tools and technologies. We have discussed big data mining challenges and issues. Finally we proposed a research direction to integrate granular computing with big data mining to help in solving these challenges.

References

- [Koliopoulos et. al , 2015] Koliopoulos AK, Yiapanis P, Tekiner F, Nenadic G, Keane J. "A Parallel Distributed Weka Framework for Big Data Mining using Spark". Big Data (BigData Congress), IEEE International Congress on. IEEE. 2015: 9–16. (2015).
- [Alkowaileet et. al , 2016] Wail Y. Alkowaileet , Sattam Alsubaiee , Michael J. Carey , Till Westmann , Yingyi Bu, "Large-scale complex analytics on semi-structured datasets using asterixDB and spark" , Proceedings of the VLDB Endowment, v.9 n.13, p.1585-1588, September, (2016) .
- [Park and Peng , 2016] K Park, L. Peng, "A Design of High-speed Big Data Query Processing System for Social Data Analysis: Using Spark SQL", International Journal of Applied Engineering Research, vol. 11, no. 14, pp. 8221-8225, (2016).
- [Naacke et. al , 2016] Hubert Naacke, Olivier Curé, Bernd Amann: SPARQL Query Processing with Apache Spark In:CS.DB 3 Nov 2016 ACM, pages 10–23, (2016).
- [Yao et. al , 2013] Yao JT, Vasilakos AV, Pedrycz W. , " Granular computing: perspectives and challenges. IEEE Trans Cybern 43(6):1977–1989. (2013) .
- [Yao , 2006] Y.Y.Yao, “Granular computing for data mining,” in Proc. SPIE Conf. Data Mining, Intrusion Detection, Inf. Assur.,Data Netw. Security,Orlando, FL , pp. 1–12 .(2006)
- [Zadeh, 1997] Zadeh, L.A. "Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic", Fuzzy Sets and Systems, 19, 111-127, (1997).
- [Lin, 1997] Lin, T.Y. "From rough sets and neighborhood systems to information granulation and computing in words" , Proceedings of European Congress on Intelligent Techniques and Soft Computing, 1602-1607, (1997).
- [Yao , 2010] Yao YY. , "Human-inspired granular computing". In: Yao JT (ed) Novel developments in granular computing: applications for advanced human reasoning and soft computation. IGI Global, Hershey, Philadelphia, pp 1–15 . (2010) .
- [Yao , 2008] Yao, Y. Y. , "The rise of granular computing". Journal of Chongqing University of Posts and Telecommunications, (Natural Science Edition). 20, 299–308. (2008).

- [Skowron et. al , 2004] Skowron, A., Jankowski, A., Dutta, S.: "Interactive granular computing". *Granul.Comput.* 1, 95–113 (2016)
- [Dean and Ghemawat, 2004] Jeffrey Dean , Sanjay Ghemawat, "MapReduce: simplified data processing on large clusters" , *Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation*, p.10-10, December 06-08, , San Francisco, CA.(2004).
- [Zaharia et. al , 2010] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in *Proc. 2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud)*, (2010).
- [Katsipoulakis et. al , 2015] Katsipoulakis NR, Tian Y, Reinwald B, Pirahesh H. "A Generic Solution to Integrate SQL and Analytics for Big Data". In: *18th International Conference on Extending Database Technology (EDBT)*; pp. 671–6. (2015).
- [Wani and Jabin, 2018] Wani M.A., Jabin S. , " Big Data: Issues, Challenges, and Techniques in Business Intelligence" . In: Aggarwal V., Bhatnagar V., Mishra D. (eds) *Big Data Analytics. Advances in Intelligent Systems and Computing*, vol 654. Springer, Singapore (2018).
- [Pal and Saini, 2014] Pal, K. & Saini, J.R.. A study of current state of work and challenges in mining big data. *International Journal of Advanced Networking* 4(7), 73-76. (2014).
- [Kumar and Kumari, 2015] Kumar, R.R. and Kumari, B. " Visualizing big data mining: Challenges, problems and opportunities<". *International Journal of Computer Science and Information Technologies* 6, (4): 3933–3937.(2015).
- [Mall and Rana, 2016] Mall, S., Rana, S., "Overview of big data and Hadoop". *Imperial Journal of Interdisciplinary Research*. Vol. 2 , Issue -5 , 1399–1406 , (2016) .
- [Sherin et. al , 2014] Sherin A, S Uma, Sarnya K, Sarnya Vani M, "Survey On Big Data Mining Platforms, Algorithms And Challenges", *International Journal of Computer Science & Engineering Technology (IJCSET)* , Vol. 5 No.9 , 854-860 (2014).

جامعة القاهرة
معهد الدراسات والبحوث الإحصائية

المؤتمر السنوى الثالث والخمسين
للإحصاء وعلوم الحاسب
وبحوث العمليات

علوم الحاسب والمعلومات

3-5 ديسمبر 2018