Cairo University
Institute of Statistical Studies and Research

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

Information Systems

3-5, Dec, 2018

Index
Information Systems

A Web Service Substitution Method based on K-means Clustering

Doaa H. Elsayed[1], Eman S. Nasr[2], Alaa El Din M. El Ghazali[3], Mervat H. Gheith[1]

Abstract

Web Service (WS) has attracted considerable attention and research to support Service Oriented Architecture (SOA). WSs execute in a highly dynamic environment, as a result, the Quality of Service (QoS) of a WS is constantly evolving; service providers may change QoS properties of WSs; existing WSs may be unavailability; new WSs may become available; variations in the infrastructure may affect the performance of the existing WSs. when the performance of WS is impaired, the failure WS is substituted with another equivalent one from a list of candidate WSs. As far as we know, almost all the existing approaches in this domain focus only on the WS substitution strategy without paying attention to reduce the time complexity of the searching algorithm. Therefore, in this paper k-means clustering is used to prune the WSs in the search space to find the best WSs for the environment changes. We implemented this method over the .NET Framework using C# programming language. A series of comparative experiments were carried out showed that the proposed method can substitute the failure WSs with another equivalent one in a time efficient manner.

Keywords: Quality of services, Web service, k-means clustering.

[1] Institute of Statistical Studies and Research, Cairo University, Cairo, Egypt

[2] Independent Researcher, Cairo, Egypt

[3] Sadat Academy for Management Sciences, Cairo, Egypt

1. Introduction

Service-Oriented Computing (SOC) is the computing paradigm that utilizes services as fundamental elements to create flexible, dynamic business processes and agile applications that span organizations and computing platforms (Li, 2014). This is achieved through Service Oriented Architecture (SOA), which is an architectural approach to develop distributed systems in heterogeneous environments that deliver an application's functionality as services that are independent of any language and platform (Hsieh & Lin, 2016; Ide & Pustejovsky, 2010). There are different implementation models and techniques for SOA but the most common is Web Services (WSs) which make interoperability easy to implement between disparate systems. WS defines as "a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically Web Services Description Language (WSDL)). Other systems interact with the WS in a manner prescribed by its description using Simple Object Access Protocol (SOAP) messages, typically conveyed using Hyper Text Transfer Protocol (HTTP) with an eXtensible Markup Language (XML) serialization in conjunction with other Web-related standards." (Sommerville, 2011). The WSs architecture is based on the interactions between three primary roles: service provider, service registry, and service consumer (D'Mello & S, 2010). These roles interact using publish, find, and bind operations (Kreger, 2015). The service provider is the business that provides access to the WS and publishes the service description in a service registry (Kreger, 2015). The service requestor finds the service description in a service registry and uses the information in the description to bind a service (Kreger, 2015; Karthikeyan & M., 2014). Service Registry is a searchable registry of service descriptions where service providers publish their service descriptions (Kreger, 2015). A logical view of the WSs architecture is shown in Fig. 1. The standard technologies for implementing WSs are WSDL, Universal Description, Discovery and Integration (UDDI), and SOAP (Sommerville, 2011).

WSDL is a standard for specifying WS interfaces. A service interface is a set of operations, each defining its input and output message. The structure of the data within a message is defined by XML Schema (Sommerville, 2011). UDDI standardizes registries that WSs can exploit to advertise themselves on the Web (Sommerville, 2011). In turn, the customer can use the registries to discover, locate and execute WSs. SOAP describes how the WSs state is affected by operation execution and suggests valid operation sequences (Sommerville, 2011).

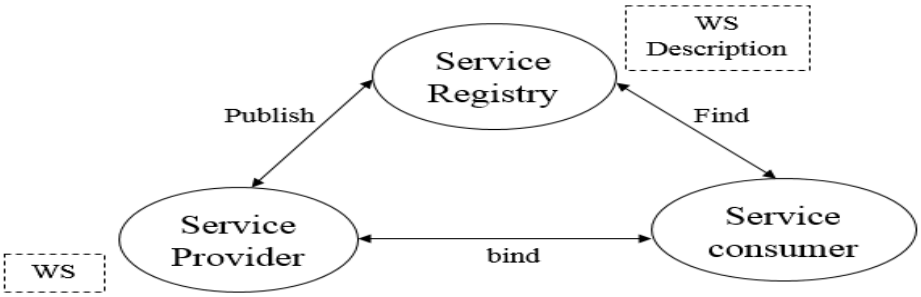


Fig. 1. WSs Architecture.

The descriptions of a WS could be divided into Functional Requirements (FR) and Non-Functional Requirements (NFR) (Rostami, et al., 2014). FR represents the functionality of a WS (i.e., execution of a WS), and NFR represents the quality of the WS, such as QoS, scalability, usability, maintainability, etc. WSs selection is based on FR and NFR. NFR have attracted great research attention in the WSs selection because of larger numbers of WSs. The selection of the best WS that satisfies the FRs and optimizes QoS requirements is called QoS-aware WSC, which is an NP-hard problem (Jatoth, et al., 2015).

WSs execute autonomously in a highly dynamic environment. Accordingly, WS is characterized by a continuous evolution; service providers may change QoS properties of WSs; existing WSs may be unavailability; new WSs may become available; variations in the infrastructure may affect the performance of the existing

WSs (Cardellini, et al., 2013). Different types of faults may happen during the execution of the composite WS. Therefore, different fault tolerance strategies should be employed to recover them (Gupta, et al., 2018). The current dynamic fault tolerance strategies include forward recovery, and backward recovery (Zhang, et al., 2018). If the failure WS can be retried, replicated, or substitute, forward recovery is allowed. If the effects produced by the failure WS need to be compensated, backward recovery is allowed. In this paper, the failure WS is substituted with another equivalent one from a list of candidate WSs. As far as we know, almost all the existing approaches in this domain focus only on the WS substitution strategy without paying attention to reduce the time complexity of the searching algorithm. Therefore, in this paper, the WSs are clustered based on QoS values by using a k-means clustering algorithm to substitute the failure WSs with another equivalent one in a time efficient manner.

The rest of this paper is organized as follows. Section 2 gives a brief review of the related literature. Section 3 presents our k-means clustering substitution method. Section 4 gives an evaluation of our method. Finally, Section 5 gives the conclusion and future work.

2. Related Work

In this section, we review the related work of WS substitution in the literature. Gupta et al. (Gupta, et al., 2018) proposed approach that is responsible for the trusted environment and complexity reduction at the component level before a fault occurs. It contains Seven modules such as UDDI registry, request collector, pre-processor, optimum service selector, web services orchestrator, fault detector, and fault manager. Once a faulty WS is detected, the fault WS will replace with available optimum. The optimum WS is WS with maximum QoS. This approach is not suitable for the dynamic list of candidate WSs. Gupta and Bhanodia (Gupta & Bhanodia, 2013) proposed a subset replacement mechanism which replaces the subset of failed

WSs with another equivalent subset. After determined equivalent subsets, equivalent subsets were ranked and the best subset among them was selected. Their mechanism was the prototype and There was not any implementation. Gupta and Bhanodia (Jayashree & Anand, 2015) proposed runtime faults detection process. Once, failure WS detected in their approach, the human intervention is required by displaying the meaningful error message to the user to understand why and where the failure has occurred. The user takes the necessary decision that resubmits the service request or substitutes the failure WS with equivalent one. Their approach is tested by using a sample WS application. Rajendran et al. (Rajendran, et al., 2017) substituted the failure WS based on the response time and the availability of the WSs. Karray et al. (Karray, et al., 2011) proposed an approach which is used aspect-oriented programming and case-based reasoning to enhance the reliability of WSs. Their approach is implemented in weather WS. Li et al. (Li, et al., 2013) also proposed a framework based on case-based reasoning. Their framework consists of a business process part, an extractor part, and a case-based reasoning part. The business process achieves the users' requirements. The extractor part extracts the FR, NFR and fault information. The faulty information is classified to the case with solution and case without the solution. Their framework is not efficient due to the case base becoming very large and it doesn't have techniques to maintain it.

Boumhamdi and Jarir (Boumhamdi & Jarir, 2012) presented the state of the art of various types of failures and recovery methods. They also proposed the architecture for detecting and dynamically recovering failure of the Business Process Execution Language (BPEL) process based on the FR of a user's preferences. This architecture has consisted of request analyzer, discovered component, constraints manager, selection manager, and orchestrator. When the failure WS occurs, the BPEL adapter is substituted the failure WS with an alternative one. Subramanian et al. (Subramanian, et al., 2008) extend BPEL with WS failure detection to enhancement it. This policy monitors the BPEL activities due to unexpected failures in WS process,

diagnoses the cause of failure and suggests a solution to this failure. Diagnoses part policy uses the database to store failure information to detect it. If failure information is not in the database, the human intervention is required. Poonguzhali et al. (Poonguzhali, et al., 2014) presented a WS substitution strategy which focuses on WS unavailability. Their approach consists of BPEL monitor, diagnoser and path substituter. Their approach alternates the routing path which contains the failure WS to the nearby router to the composer. Their mechanism was the prototype and There was not any implementation.

3. K-means Clustering Substitution Method

To efficiently substitute a failure WS, this paper applies a K-means clustering algorithm to build WS clusters. The WSs in the same cluster have nearly QoS attributes. Therefore, the efficiency of WS substitution can be significantly improved by prune the WSs in the search space to the candidate WSs in the same cluster of failure WS. Every WS has multiple QoS attributes which are classified to positive and negative attributes. The values of negative attributes need to be minimized, such as response time and cost. When the value of negative attributes is smaller, the performance is better. The values of positive attributes need to be maximized, such as availability and reliability. When the value of positive attributes is bigger, the performance is better. Multiple Criteria Decision Making (MCDM) is used to normalize all QoS attributes to the same scale in a range from 0 to 1 to allow a standard measurement of QoS level for a single WS, especially with the multiple criteria. The MCDM of negative and positive QoS attributes is given by (1) (Shehu, et al., 2014).

$$\begin{Bmatrix} \sum_{j=1}^{r} \frac{(Q_j^{max} - Q_{ij})}{(Q_j^{max} - Q_j^{min})} * w_j \\ \sum_{j=1}^{r} \frac{(Q_{ij} - Q_j^{min})}{(Q_j^{max} - Q_j^{\min})} * w_j \\ \sum_{j=1}^{r} 1 * w_j \end{Bmatrix} \begin{matrix} if Q_j^{max} - Q_j^{min} \neq 0 \\ if Q_j^{max} - Q_j^{min} \neq 0 \\ if Q_j^{max} - Q_j^{min} = 0 \end{matrix} \quad (1)$$

Equation 1 assumes that each WS has *r* service quality criteria. A matrix *Q* is built, in which each row Q_j corresponds to a WS i, while each column corresponds to a quality dimension. And Q_j^{max} and$Q_j^{\min}$ are the maximal value and minimal value of a quality criterion in the matrix *Q* respectively. $w_j \epsilon$ [0; 1] represents the weight criterion j of which the values are provided by the users based on their own preferences.

K-means clustering is clustered candidate WSs into clusters based on MCDM value. The input of the k-means clustering algorithm is the number of clusters and the MCDM value of all candidate WSs. The k-means clustering steps are (Tan, et al., 2006):

A. Select random *K* points as the initial centroid.

B. Calculate the weighted Euclid distance between each candidate WSs and the centroid. The dataset is assigned to the cluster according to the minimum distance *d* formulated as in (2) (Tan, et al., 2006):

$$d = \sqrt{(a - \bar{a})^2} \quad (2)$$

where a denotes the MCDM value of the candidate WSs and $\bar{a}$ denotes the value of the centroid.

C. When all the candidate WSs are assigned to their respective clusters, recomputed the average of the centroid and get a new centroid of the *K* clusters.

D. Repeat steps B and C until the centroids don't change.

As shown in Fig. 2, In the case of emerging new WS to candidate WSs, the MCDM of new WS is calculated and this WS is assigned to the corresponding WS cluster

based on MCDM value. In the case of variation in WS performance, the WS is removed from the WS cluster and then assign it into corresponding WS cluster based on new MCDM value. In the case of unavailability of WS, this WS is removed from assign cluster. When the failure WS occurs, it is a substitute with alternative WS in the same cluster because this cluster contains the nearest WS to the failure WS. The process of the WS Substitution strategy is illustrated:

1. The MCDM Value and the cluster number of the failing WS are retrieved.
2. The candidate WSs in the same cluster of the failing WS are retrieved.
3. Calculate the substitution Coefficient to quantify the similarity degree between the failure WS and candidate WSs in the same cluster by using (3):

$$Sc(Q_i.Q_{j)} = \frac{F_{Q_j}}{F_{Q_J} - F_{Q_i}} \qquad (3)$$

   Where $Sc(Q_i.Q_{j)}$ represents substitution Coefficient between unavailable WS j and candidate WS i. Q_i represents candidate WS. Q_J represents the unavailable WS. F_{Q_j} represents the fitness value of unavailable WS and F_{Q_i} represent the fitness value of candidate WS i.
4. The candidate Web service with the higher substitution Coefficient will be selected.

5. Evaluation

The experiment background is designed as follows. We consider three QoS attributes of services, namely, service cost, response time, and service reliability. The weight of service cost, response time and service reliability are 0.5, 0.2 and 0.3. The experiment's environment is a Dell Laptop with an Intel Core i7 at 2.50 GHz, 8 GB RAM, running Microsoft Windows 10. The algorithm is implemented in Visual Studio 2015 using C# language. The QoS of the candidate Web services and the centroid of the clusters are saved in SQL server 2014. The execution time measured by MilliSeconds (MS).

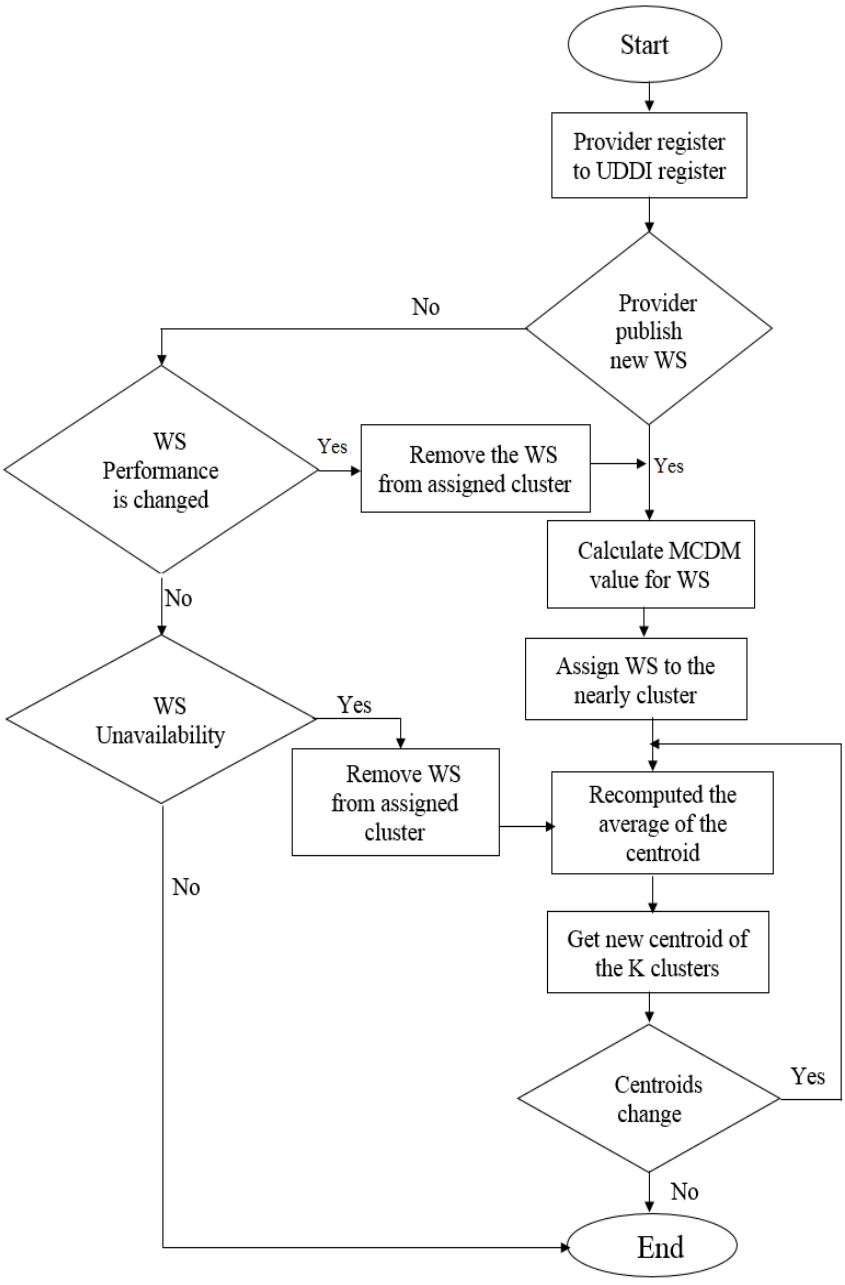


Fig. 2. The Flowchart of applying K-means Cluster to Candidate WSs.

Fig. 3 illustrates the computation time for substituting the failure WS to another equivalent WS from the candidate WSs. In this experimental, the candidate WSs is varied from 500 to 3000 WSs. The number of clusters is 5 clusters. In 3000 candidate WSs, in the case of clustering candidate WSs the computation time for substituting the failure WS to another equivalent WS is 380MS while the computation time for substituting the failure WS to another equivalent WS without clustering is 675 MS. Hence it could be deduced that the computation time for the WS substitution in the case of clustering candidate WSs by using k-means clustering is much more effective.

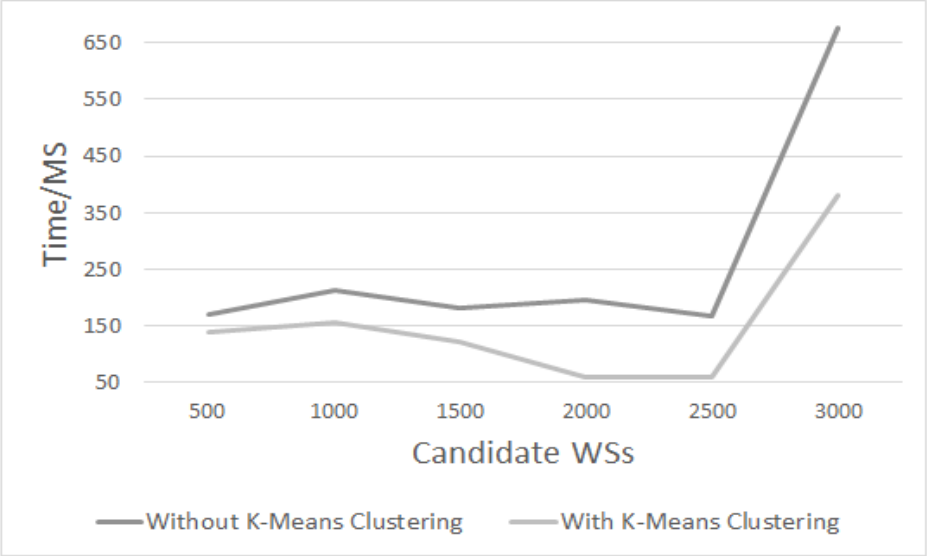


Fig. 3. The Computation Time for Substitute the failure WS to another equivalent WS with a different Number of the Candidate WSs.

Fig. 3 illustrates the computation time for substituting the failure WS to another equivalent WS from the candidate WSs in the different number of clusters. In this experimental, the number of clusters is varied from 5 to 50 clusters and the number of candidates WSs are 5000 WSs. In the case of clustering candidate WSs in 5 clusters the computation time for WS substitution is 207 MS. Then the curve progressively increased in the case of clustering candidate WSs in 10 and 15 clusters. After that,

the curve went down in the case of clustering candidate WSs in 20, 25 and 30 clusters. Then the curve is stable in the case of clustering candidate WSs from 30 to 50 clusters. Hence it could be deduced that the best number of clusters in this experimental 30 clusters.

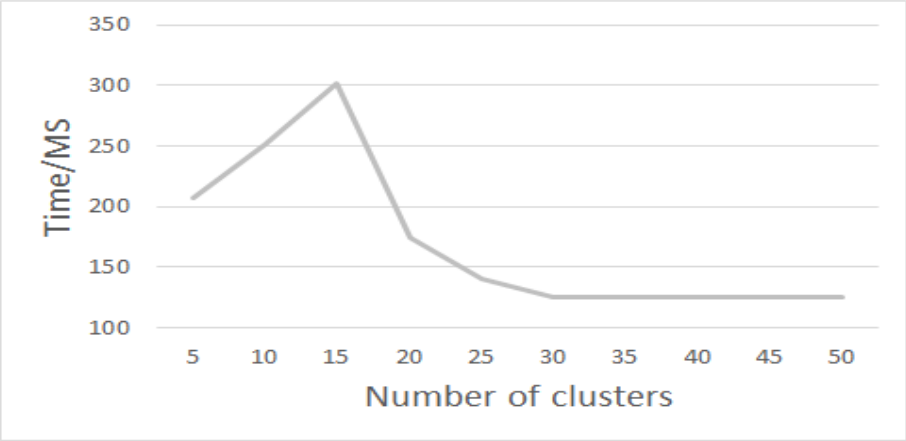


Fig. 4. The Computation Time for Substitute the failure WS to another equivalent one with a different Number of the clusters.

4. Conclusion and Future Work

This paper proposed WS clustering method based on QoS attributes to support dynamic QoS requirement with automatic adaptation with respect to change nature of QoS of WSs. K-means clustering technique is utilized to clustering candidate WSs. In the case of QoS changes, the failure WS sustains by decreasing the search space to the WSs in the same cluster of the failing WSs. Experiment results show that our method can substitute the failure WSs with another equivalent one in a time efficient manner. In our future work, we intend to integrate this method to monitoring framework to get run-time QoS matrices of the WSs. Eventually, we would like to evaluate our method by studying the usability of our method in real life service-based systems.

References

Boumhamdi, K. & Jarir, Z. (2012). "An Approach to Support Monitoring and Recovery of BPEL Processes at Runtime". International Journal of Computer Applications. Vol. 43, No.2. pp. 34 - 41.

Cardellini, V. et al. (2013). "MOSES: a Platform for Experimenting with QoS-driven Self-adaptation Policies for Service Oriented Systems". In: R. d. Lemos, D. Garlan, C. Ghezzi & H. Giese, eds. Software Engineering for Self-Adaptive Systems III. Assurances. Lecture Notes in Computer Science. Germany: Springer, Cham, pp. 409-433.

D'Mello, Demian. Antony; S, Ananthanarayana, V. (2010). "Dynamic Web Service Composition Based on Operation Flow Semantics". International Journal of Computer Applications, February. Vol. 1.No. 26. pp. 1-9.

Gupta, R., Kamal, R. & Suman, U. (2018)." A QoS-Supported Approach using Fault Detection and Tolerance for achieving Reliability in Dynamic Orchestration of Web Services". International Journal of Information Technology Vol. 10, No. 1. pp. 71–81.

Gupta, S. & Bhanodia, P. (2013)." A Fault Tolerant Mechanism for Composition of Web Services using Subset Replacement". International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, No. 8, pp. 3080-3085.

Hsieh , F.-S. & Lin, J.-B. (2016). "A Self-adaptation Scheme for Workflow Management in Multi-agent Systems". Journal of Intelligent Manufacturing. Vol. 27, No. 1, pp. 131–148.

Ide, N. & Pustejovsky, J. (2010). "What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability for Language Technology". Hong Kong. In Proceedings of the 2nd International Conference on Global Interoperability for Language Resources (ICGL).

Jatoth, C., Gangadharan, G. & Buyya, R. (2015). "Computational Intelligence based QoS-aware Web Service Composition: A Systematic Literature Review". IEEE Transactions on Services Computing. Vol. 10, No.3. pp. 475-492.

Jayashree, K. & Anand, S. (2015). "Run Time Fault Detection System for Web Service Composition and Execution". Smart Computing Review. Vol. 5. No. 5, pp. 469-482.

Karray, M.-H., Ghedira, C. & Maamar, Z. (2011). "Towards a Self-Healing Approach to Sustain Web Services Reliability".In Proceedings of the 2011 Workshops of International Conference on Advanced Information Networking and Applications. pp. 268-272.

Karthikeyan, J. & Kumar Suresh, M. (2014). "Monitoring QoS parameters of composed web services". In Proceedings of the International Conference on Information Communication And Embedded Systems (ICICES). pp. 24 - 29.

Kreger, H., 2015. Web Services Conceptual Architecture (WSCA1.0). [Online].

Li, G. et al. (2013). "A Self-healing Framework for QoS-Aware Web Service Composition via Case-Based Reasoning". In: Web Technologies and Applications. APWeb 2013. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 654-661.

Li, W. (2014). "Towards a Resilient Service Oriented Computing based on Ad-hoc Web service compositions in Dynamic Environments(Doctoral Dissertation)". Institut d'Optique Graduate School.

Poonguzhali, S., JerlinRubini, L. & Divya, S. (2014). "A Self-Healing Approach for Service Unavailability in Dynamic Web Service Composition". In Proceedings on the International Journal of Computer Science and Information Technologies, pp. 4381-4383.

Rajendran, V., Chua, F.-F. & Chan, G.-Y. (2017). "Self-Healing in Dynamic Web Service Composition". In Proceedings of the 5th International Conference on Future Internet of Things and Cloud. pp. 206-211.

Rostami, N. H., Kheirkhah, E. & Jalali, M. (2014). "An Optimized Semantic Web Service Composition Method Based on Clustering and Ant Colony Algorithm". International Journal of Web & Semantic Technology. January. Vol. 5. No. 1, pp. 1-8.

Shehu, U., Epiphaniou, G. & Safdar, G. (2014). "A Survey of QoS-aware Web Service Composition Techniques". International Journal of Computer Applications. Vol. 89. No. 12, p. 11.

Sommerville, I. (2011). "Software Engineering (9th Edition)". Pearson.

Subramanian, S. et al. (2008). "On the Enhancement of BPEL Engines for Self-Healing Composite Web Services". Turku, Finland. In Proceedings of the 2008 International Symposium on Applications and the Internet. pp. 33-39.

Tan, P.-N., Steinbach, M. & Kumar, V. (2006). "Introduction to Data Mining". Pearson.

Zhang, J. et al. (2018). "Overview on Fault Tolerance Strategies of Composite Service in Service Computing". Wireless Communications and Mobile Computing. p. 8.

A Proposed Plan for Crowdsourcing as a Requirements Elicitation Method for eLearning Systems

Nancy M. Rizk[1] Eman S. Nasr[2] Mervat H. Gheith[1]

***Abstract*:** eLearning is generally defined as the electronic access to learning resources, anywhere and anytime. eLearning systems are diverse in terms of size and nature, they became a very important part of teaching, both as web-based systems for online education, or as supportive tools for physical on-campus study. With the diversity of people using eLearning systems, there is a need for continuous improvement, and better understanding for the stakeholders' requirements by the software development teams. Requirements elicitation is an early activity in requirements engineering process that is concerned with discovering and determining needs of stakeholders. Crowdsourcing is the use of contributions of people in order to complete a service or task. The paper proposes the use of crowdsourcing approach in the requirements elicitation phase when developing eLearning systems. This paper provides a proposed plan to review on eLearning systems, requirements elicitation and crowdsourcing in attempt to propose a new method based on crowdsourcing approach that helps to support the elicitation of requirements from a large number of stakeholders for enhancing or developing a new eLearning software project.

Keywords: Requirements engineering; RE; requirements elicitation; electronic Learning; eLearning; crowdsourcing.

1. INTRODUCTION

eLearning is one of the most important concepts developed after the rise of internet. It ranges from sharing simple files, to getting a certain degree over the internet. eLearning systems is used by the educational institutions as a supportive tool for collaboration between users, providing more facilities e.g. submitting assignments, taking quizzes, and manage learning materials. Although there is no one definition for eLearning, but generally it can be defined as the use of technology in the delivery of education or training (Australian Flexible Learning framework, 2007), or the online access to learning resources, anywhere and anytime (Epignosis LLC, 2014), (Holmes & Gardner, 2006). eLearning systems are nowadays integral parts of some commercial or governmental organizations, as linking the gap between learning and work can give great benefits. It can help to get cost-effective trainings, and employees can share experiences, so that knowledge can be leveraged in the business.

eLearning systems have special characteristics or properties, one is the participants' diversity; the large number of stakeholders lead to diversity of users from different aspects e.g. background, culture, regulations and geographical (Alharthi, Spichkova, & Hamilton, 2015), (Goldsworthy & Rankine, 2009), (Piccioli). Collaboration is another characteristic because improving social interaction in eLearning system leads to more satisfaction for the learning process [26]. One last characteristic is the large number of stakeholders; evolution of web technologies led to great increase in the number of people involved in eLS. Users are learners, instructors, and management.

[1]Institute of Statistical Studies and Research (ISSR), Cairo University, Egypt
[2]Independent Researcher

With the great number and diversity of the users participating in eLearning systems, there is a need for constant improvement and better understanding to the stakeholders' requirements in the software requirements determination phase, in order to reach for a successful eLearning system.

RE was defined by DOD in 1991 to be the phase that "involves all life-cycle activities devoted to identification of user requirements, analysis of the requirements to derive additional requirements, documentation of the requirements as a specification, and validation of the documented requirements against user needs, as well as processes that support these activities" (Hull, Jackson, & Dick, 2011). RE activities or processes according to Sommerville (Sommerville, 2011) are: requirements elicitation, requirements analysis, requirements specification and requirements validation.

Requirements elicitation according to Zowghi and Coulin (Zowghi & Coulin, 2005) is "the process of seeking, uncovering, acquiring, and elaborating requirements for computer based systems. This activity is very important for the success of the software development project, where detecting errors from the beginning in this stage can save time and money (Standish Group International, 2013). According to Standish group report (Standish Group International, 2013) user involvements is one of the top success factors in any software development project and this is beginning to happen in requirements elicitation phase that has a social nature because most of its activities depend on people (Fernandes, Duarte, Ribeiro, Farinha, Pereiraa, & Mira da Silva, 2012). Examples of traditional requirements elicitation approaches are; interviews, group work, ethnography. The large number of stakeholders or end users made difficulty in traditional activities of requirements elicitation, so there is a great need for newer and up to date activities that take benefits from the existing concepts of sharing in web 2.0 and social networking to enhance the elicitation phase. One of these new activities that we want to use in the requirements elicitation is crowdsourcing.

Crowdsourcing is an emerging business model where tasks are accomplished by the general public; the crowd (Hosseini, Phalp, Taylor, & Ali, 2014). It is defined in Merriam Webster as "the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers" (Merriam Webster). The term first use started in 2006 by Jeff Howe in his article "The rise of crowdsourcing" (Howe, 2006). The revolution that happened in internet after introducing Web 2.0 and the great success of the concept of read and write web and its applications like; Wikipedia, or Facebook and Twitter, crowdsourcing became a powerful tool for different disciplines e.g. Marketing, Information systems. Hosseini et al. (Hosseini, Shahri, & Dalpiaz, Configuring Crowdsourcing for Requirements Elicitation, 2015) mentioned that crowdsourcing in requirements elicitation has the potential to increase the quality and comprehensiveness of requirements elicitation. Crowdsourcing allow great access of software development team of a wide diversity of potential and actual stakeholders. This paper presents a study on how crowdsourcing concept can be useful in the area of requirements elicitation for eLearning systems.

The rest of this paper is organized as follows Section 2 analyzes previous work in the area under research. Section 3 provides the proposed Plan of the research, and finally, section 4 discusses conclusion and future work.

2. PREVIOUS WORK

We searched in previous work part in three areas: RE for eLearning systems, crowdsourcing methods of RE, and supportive crowdsourcing tools of RE. Aiming to investigate if there is a special process of RE to eLearning systems, analysis of the previous work of crowdsourcing methods used in RE, and studying the crowdsourcing tools completed to support the RE processes.

A. *RE for eLearning systems*

Alharthi et al. (Alharthi, Spichkova, & Hamilton, 2015) raised some questions that will be answered in future publications about what is specific for e-learning systems in RE, how RE processes can be improved for e-learning systems, and how to deal with the diversity aspects when developing global e-learning system. Then it will develop a methodology for the RE of e-learning systems.

B. *Crowdsourcing methods of RE*

Fernandez et al. (Fernandes, Duarte, Ribeiro, Farinha, Pereiraa, & Mira da Silva, 2012) presented iThink game based tool for requirements elicitation. It is a tool to be used by stakeholders to get information about existing requirements or new requirements based on a gaming approach with points and rewards for the winner. The problem in this tool is that it only works for a limited number of stakeholders, and is not capable of getting requirements from end users or a crowd. The work is evaluated using case studies and their future work will be concerned with evaluating the tool using more case studies.

Senjiders (Snijders, Ozum, Brinkkemper, & Dalpiaz, 2015) proposed Crowd Centric Requirement Engineering (CCRE); it is a method based on gamification and crowdsourcing to involve users' crowd in requirements elicitation, negotiation and prioritization, and it gives rewards for participation. The problems in this method are the need of professional management to moderate the requirements, high expectations from stakeholders, and ambiguous requirements. It was evaluated by applying the tool on case studies, and they propose future work to apply it on larger projects and for a longer time period. It was also evaluated by interviewing three experts and studying their opinions.

Lim et al. (Lim & Ncube, 2013) used the social networking to build a network of stakeholders and recommended stakeholders. The study also applied social networks analysis measure to identify and prioritize stakeholders and requirements.

Seyff et al. (Seyff, Todoran, Caluser, Singer, & Glinz, 2015) performed three studies based on brainstorming approach on Facebook where students involved in order to participate in requirements elicitation, negotiation, and prioritization. The three studies showed that social networks sites can be used to help in requirements elicitation on large scale, getting end users need and new ideas but need to be moderated well. Future work is presented in number of points some related to the method of evaluation and the need to apply on more and larger case studies, another one related to applying the approach on different social networks sites like LinkedIn. One last point is studying the motivation of end users to collaborate in requirements elicitation process; they already started investigation and found some end users collaborate because of their duty while others because of their interest in the topic, supporting a friend, curiosity.

Hosseini et al. (Hosseini, Shahri, & Dalpiaz, Configuring Crowdsourcing for Requirements Elicitation, 2015) presented an empirical study to advocate the use of crowdsourcing in requirements elicitation, it involved 14 participants in two focus groups of users and developers, then an online expert survey was conducted involving 34

participants from the RE community. It discussed some challenges related to largeness, diversity, anonymity, volunteering and feedback.

Groen et al. (Groen, Doerr, & Adam, 2015) performed a research preview for crowd-based RE, it discusses two main points. First, the solution developed or suggested to overcome the deficit in traditional RE. Second how can user feedback be categorized in relevant dimensions. Future work is the following, and we are interested in point number one which is; "How to solve the reticence or concerns of stakeholders regarding the provision of feedback through the use of suitable motivational instruments and appropriate levels of intrusiveness?"

C. Supportive crowdsourcing tools of RE

This section discusses the existing tools that are used to get, and analyze the users' feedback. In order to use them in collecting requirements, these tools can be commercial tools or proposed tools under research.

CrowdREquire was first introduced in a technical report by Adepetu et al. (Adepetu, Ahmed, & Al Ab, 2012). The report describes CrowdREquire as a platform that supports crowdsourcing RE. It helps companies and individuals to find the best requirements specification for their proposed tasks and projects. It also has a communication tool to connect RE professionals with the companies that request their services. CrowdREquire provided a list of requirements specification written in consistent level of detail. The evaluation of CrowdREquire was done and checked for different aspects however the method of evaluation was not mentioned clearly. Also, Srivastava and Sharma (Srivastava & Sharma, 2015) have proposed a crowdsourcing-based solution to a case study on MyERP software to extract software requirements across from various and different geographies crowd. The solution succeeded in getting diverse opinions & requirements, and to lower the cost of it.

OpenPropsal is proposed by Rashid et al. (Rashid, Wiesenberger, Mede, & Baumann, 2008) with the purpose of helping software engineers in the requirements management process. It allows users to annotate their feature requests, error reports or enhancement requests and directly send them to the requirements management, then track progress on their requirements. The requirements analysts take part in the discussion of the proposals, prioritize, them and decide on what proposal gets implemented and what doesn't, and also propose their own ideas and discusses with the other stakeholders. The software engineer understands the users' proposals and to guarantee their correct implementation. He can submit his own proposal specifications, participate in discussions to what is technically possible, and is responsible for implementing the proposals that have been decided on. OpenProposal was evaluated in different ways. First, using usability test to ensure the usability of annotation tool, in this test an interview questionnaire and observation were conducted to gather information about the system. Second, using case studies in the company and the university were the system has been established.

Requirements Bazar is a web based social software tool for social RE presented by Renzel et al. (Renzel, Behrendt, Klamma, & Jarke, 2013), and now it is a real project that can be used (Requirements Bazaar). It focuses on four aspects: a) requirements specification which is an aggregation of basic metadata, artifacts and traceable means for social participation, b) a workflow for co-creation which involves many processes that integrate communities into the entire service development process, c) workspace integration, this is to lower entry barriers, to integrate with end-user and developer workspaces, and d) personalizable requirements prioritization, supports requirements prioritization with a modular extensible requirements ranking framework. The evaluation

of Requirements Bazaar was conducted using short term study of 2 groups of service providers and requesters followed by questionnaire assessing five dimensions (System Quality, Information Quality, User satisfaction, Individual Impact and Community Impact). The system also evaluated by user experiences from long term productive use, they collect the votes and feedbacks of some aspects of the system (Renzel & Klamma, Requirements Bazaar: Open-Source Large Scale Social Requirements Engineering in the Long Tail, 2014).

Natural language-oriented tools are used in crowdsourcing in two ways. First are ontologies to support elicitation and refinements. The second direction is to use text-mining tools on structured sets of requirements documents to cluster and categorize similar requirements from different sources and then use stakeholder ownership details to prioritize requirements (Sutcliffe & Sawyer, 2013).

Guzman and Maalej presented an automated approach using sentiment analysis to analyze the users' reviews of the mobile applications in applications stores in order to get requirements from the users' feedback that help in requirements elicitation phase (Guzman & Maalej, 2014). The approach was evaluated by comparing the results of the proposed systems with manual results performed by graduate student with certain skills and handed guide to clarify their work.

Hosseini et al. have proposed CRAFT, which is technique that utilizes the crowd power to enrich text mining by allowing the crowd to categorize and annotate feedback through a context menu. This, in turn, helps in better identifying user requirements within forums feedback (Hosseini, Groen, Shahri, & Ali, 2017).

3. PROPOSED PLAN

Due to what we mentioned about the growing of web 2.0 concept and its implications on the size of the system users, the limitations of the current and traditional approaches of requirements elicitation, the statistics of the low success rates of software development projects, the special characteristics of the eLearning systems and the research gap in the area of RE for eLearning systems; as discussed previously in the previous work section. The researchers are motivated to find a new method of requirements elicitation for eLearning systems using the crowdsourcing method.

The objective of this research is to propose a new crowdsourcing Plan that helps to support the elicitation of requirements from a large number of stakeholders for enhancing or developing a new eLearning software project. In order to achieve this objective we will follow the following steps as illustrated in figure 1.

1. Surveying the literature for crowdsourcing in RE.
2. Surveying the literature for RE for eLearning systems in particular.
3. Identifying the special properties that characterize eLearning systems.
4. Comparing the different researches found and identifying the current unsolved problems
5. Proposing a social networking/crowdsourcing based method for the gathering of requirements for eLearning systems
6. Evaluating the proposed method.
7. Enhancing the method based on the results from the initial evaluation.

The research plan will start with surveying the literature using the systematic literature review method (SLR) to identify, analyze and interpret all available evidence (Wohlin, Host, Regnell, Runeson, Ohlsson, & Wesslen, 2000) related to the crowdsourcing concept in eliciting requirements of an e-Learning software projects. Discuss the requirements elicitation aspects of eLearning systems. A social networking based tool/site or more is going to be used to facilitate the gathering of requirements as

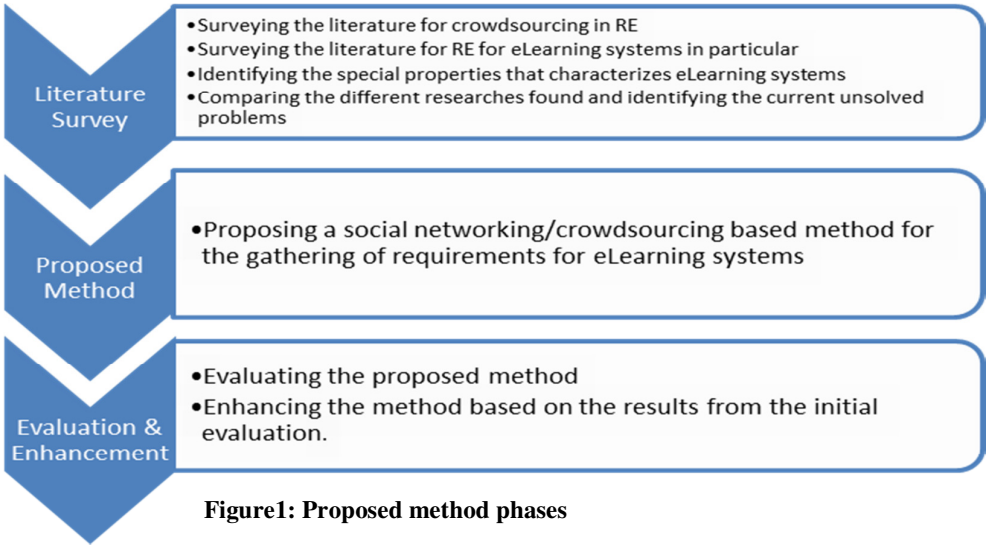


Figure1: Proposed method phases

illustrated in figure 2 that shows the architecture of the proposed approach. The study is going to be evaluated through case studies using university students, then statistical analysis methods will be used for the interpretation of the findings, and finally reporting the findings.

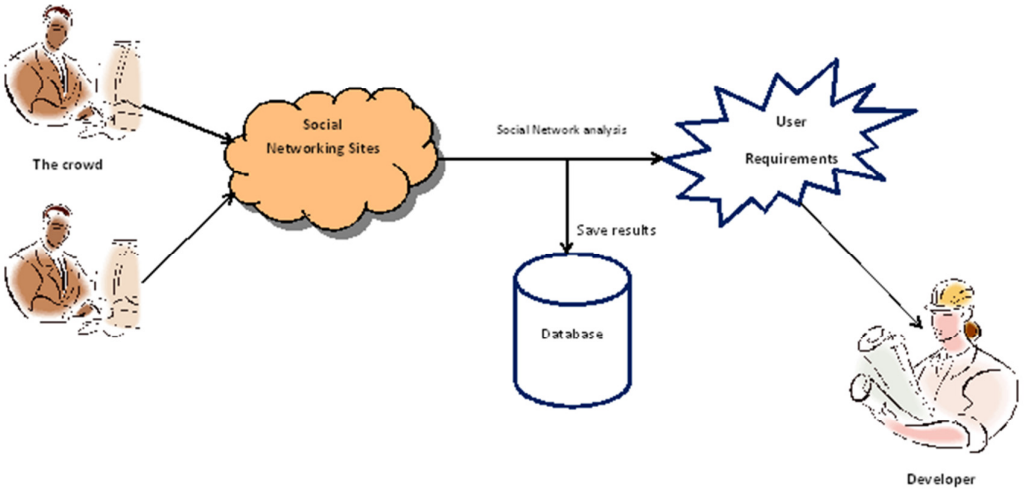


Figure 2: Proposed System Architecture

4. CONCLUSION

eLearning systems are special type of systems, they have special characteristics that should be taken into consideration in software development process; the number of stakeholders, the collaboration nature of eLearning systems, and the diversity of eLearning systems' users. Also, there is no special approach of RE for eLearning systems, because of these factors the researchers are motivated to find a new method of requirements elicitation for eLearning systems. This method can take the power of the social web and web 2.0 technologies to better understand the stakeholders' requirements. The method will be based on the crowdsourcing concept to suite what special about eLearning systems. The paper proposed a plan for the new method to follow in future to reach for the required results. For future work, better understanding and analysis for the problem domain and an implementation for the proposed method will be provided.

REFERENCES

(n.d.). Retrieved 07 31, 2015, from Merriam Webster: http://www.merriam-webster.com/dictionary/crowdsourcing

(n.d.). Retrieved from Requirements Bazaar: https://requirements-bazaar.org/info#/explore

Adepetu, A., Ahmed, K., & Al Ab, Y. (2012). *CrowdREquire: A Requirements Engineering Crowdsourcing Platform.* Technical Report, Association for the Advancement of Artificial Intelligence.

Alharthi, A. D., Spichkova, M., & Hamilton, M. (2015). Requirements Engineering Aspects of ELearning Systems. *24th Australasian Software Engineering Conference.* Adelaide, SA, Australia.

Australian Flexible Learning framework. (2007). *Practical Guide to E-learning for Industry.*

Epignosis LLC. (2014, Jan). E-learning: Concepts, Trends, and Applications.

Fernandes, J., Duarte, D., Ribeiro, C., Farinha, C., Pereiraa, J. M., & Mira da Silva, M. (2012). iThink: A Game-Based Approach Towards Improving Collaboration and Participation in Requirement Elicitation. *4th International Conference on Games and Virtual Worlds for Serious Applications* , (pp. 66-77).

Goldsworthy , K., & Rankine, L. (2009). Identifying the characteristics of e-learning environments used to support large units. *Proceedings of Ascilite Auckland.* Auckland.

Groen, E. C., Doerr, J., & Adam, S. (2015). Towards Crowd-Based Requirements Engineering A Research Preview. In S. A. Fricker, & K. Schneider (Eds.), *Requirements Engineering: Foundation for Software Quality* (pp. 247-253). Springer International Publishing.

Guzman, E., & Maalej, W. (2014). How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews. *Proceeding of Requirements Engineering Conference (RE), 2014 IEEE 22nd International*, (pp. 153-162). Karlskrona, Sweden.

Holmes, B., & Gardner, J. (2006). *E-learning: Concepts and Practice,.*

Hosseini, M., Groen, E. C., Shahri, A., & Ali, R. (2017). CRAFT: A Crowd-Annotated Feedback Technique. *IEEE 25th International Requirements Engineering Conference Workshops.* Lisbon, Portugal.

Hosseini, M., Phalp, K., Taylor, J., & Ali, R. (2014). The Four Pillars of Crowdsourcing: a Reference Model. *Proceedings of IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*, (pp. 1-12). Marrakech.

Hosseini, M., Shahri, A., & Dalpiaz, F. (2015). Configuring Crowdsourcing for Requirements Elicitation. *In Proceeding f The IEEE Ninth International Conference on Research Challenges in Information Science.* Greece.

Howe, J. (2006, June). Retrieved 07 31, 2015, from Wired: http://archive.wired.com/wired/archive/14.06/crowds.html?pg=4&topic=crowds&topic_set=

Hull, E., Jackson, K., & Dick, J. (2011). *Requirements Engineering* (3 ed.). Springer-Verlag London.

Lim, S. L., & Ncube, C. (2013). Social Networks and Crowdsourcing for Stakeholder Analysis in System of Systems Projects. *Proceeding of the 2013 8th International Conference on System of Systems Engineering*, (pp. 13-18). Hawaii.

Piccioli, V. (n.d.). Retrieved April 2016, from Docebo: https://www.docebo.com/2015/04/23/8-types-elearning-students/

Rashid, A., Wiesenberger, J., Mede, D., & Baumann, J. (2008). Bringing Developers and Users closer together:The OpenProposal story. *Proceedings of the Primium Subconference at the Multikonferenz Wirtschaftsinformatik*, (pp. 9-26).

Renzel, D., & Klamma, R. (2014, September). Requirements Bazaar: Open-Source Large Scale Social Requirements Engineering in the Long Tail. *IEEE Computer Society Special Technical Community on Social Networking E-Letter, 2*(3).

Renzel, D., Behrendt, M., Klamma, R., & Jarke, M. (2013). Requirements Bazaar: Social Requirements Engineering for Community-Driven Innovation. *Proceeding of Requirements Engineering Conference (RE), 2013 21st IEEE International*, (pp. 326-327). Rio de Janeiro.

Seyff, N., Todoran, I., Caluser, K., Singer, L., & Glinz, M. (2015, April). Using popular social network sites to support requirements elicitation, prioritization and negotiation. *Journal of Internet Services and Applications.*

Snijders, R., Ozum, A., Brinkkemper, S., & Dalpiaz, F. (2015). *Crowd-Centric Requirements Engineering: A Method Based On Crowdsourcing And Gamification.* Technical Report, Utrecht University, Utrecht, The Netherlands.

Sommerville, I. (2011). *Software Engineering* (9th ed.). USA: Pearson.

Srivastava, P. K., & Sharma, R. (2015). Crowdsourcing to Elicit Requirements for MyERP Application. *Proceeding of 1st International Workshop on Crowd-Based Requirements Engineering (CrowdRE)*, (pp. 31-35). Ottawa.

Standish Group International. (2013). *CHAOS manifesto* .

Sutcliffe, A., & Sawyer, P. (2013). Requirements elicitation: Towards the unknown unknowns. *Proceeding of Requirements Engineering Conference (RE), 2013 21st IEEE International*, (pp. 92-104). Rio de Janeiro.

Wohlin, C., Host, M., Regnell, B., Runeson, P., Ohlsson, M. C., & Wesslen, A. (2000). *Experimentation in Software Engineering.* Springer.

Zowghi , D., & Coulin, C. (2005). Requirements Elicitation: A Survey of Techniques,Approaches, and Tools. In A. Aurum , & C. Wohlin, *Engineering and Managing Software Requirements* (pp. 19-46). Berlin: Springer.

الملخص.
التعليم الإلكتروني هو الوصول إلى مصادر التعلم إلكترونيا في أي مكان وزمان. تتنوع نظم التعليم الإلكتروني من حيث الحجم والطبيعة، حيث انها أصبحت جزءًا مهمًا جدًا من التدريس كنظم للتعليم عبر شبكة الإنترنت ، أو كأدوات داعمة للدراسة الواقعية داخل الحرم الجامعي. مع تنوع الأشخاص الذين يستخدمون أنظمة التعلم الإلكتروني ، هناك حاجة للتحسين المستمر ، وفهم أفضل لمتطلبات أصحاب المصلحة من قبل فرق تطوير البرامج. استخراج متطلبات نظم المعلومات هو نشاط مبكر في عملية هندسة المتطلبات التي تهتم باكتشاف وتحديد احتياجات أصحاب المصلحة. مصادر الحشد هو استخدام مساهمات الأشخاص لإكمال خدمة أو مهمة. يقترح البحث استخدام نهج مصادر الحشد في مرحلة استخراج المتطلبات عند تطوير أنظمة التعلم الإلكتروني. يقدم هذا البحث مراجعة حول أنظمة التعلم الإلكتروني ، واستخراج المتطلبات ومصادر الحشد في محاولة لاقتراح طريقة جديدة تقوم على نهج مصادر الحشد الذي يساعد على دعم استحضار المتطلبات من عدد كبير من أصحاب المصلحة لتعزيز أو تطوير مشروع جديد لنظام التعليم الإلكتروني.

Extracting Topics from Educational Material Based on N-Gram and Word Co-occurrence

Mohammed Badawy [1]* Mahmood A. Mahmood [1] A. A. Abd El-Aziz [1, 2] Hesham A. Hefny[3]

Abstract:

The traditional way to extract topics from any educational material takes a lot of effort and time, especially when a huge amount of material was prepared. The proposed solution to extract topics from any educational material is based on n-gram and term co-occurrences, which calculates the keywords of the course specifications that appears in educational material. In addition to applying filter standards to improve the efficiency. In the proposed methodology, we focused on the educational material because the manual approach for selecting topics is not simple and needs more effort, time, and experience for selecting the appropriate topics for each course. To ensure the effectivness of the proposed approach, we have selected two courses in two institutions to test our proposed approach. Then we compared the extracted topics using our proposed approach with the selected topics using the manual extraction approach.

Keywords: Automatic Topic Selection; Information Education System; Text Mining; N-gram; Term Frequency; Tokenization.

1. Introduction

Recent studies indicated that more than 80% of the data in the digital space are composed of unstructured or semi-structured data [Talib, R et al., 2016]. These data are loaded with texts, and they exist nearly everywhere like in electronic documents, online web pages, and social media. Figure 1 shows the comparison between the increment of unstructured, semi-structured data and structured data by years.

[1] Dept. of Information Systems and Technology, Institute of Statistical Studies and Research, Cairo University.
[2] Dept. of Information System, College of Computer and Information Sciences, Jouf University, KSA.
[3] Dept. of Computer Sciences, Institute of Statistical Studies and Research, Cairo University.
mbadawy@live.com, Mahmoodissr@cu.edu.eg, a.ahmed@cu.edu.eg, hehefny@ieee.org

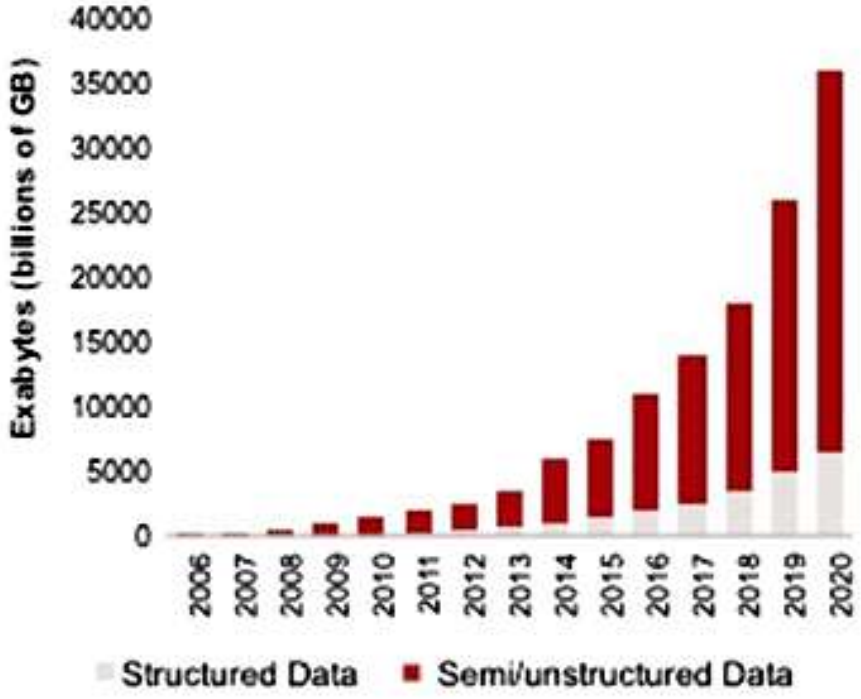


Figure 1: Growth of Data by Year [2]

Keyphrase extraction is interested in the selection of a set of phrases from a document that summarizes the main topics discussed in that document. The automatic keyphrase extraction is an essential task in digital content management because it can be used for document indexing, which enables calculating semantic resemblance between documents, also improves browsing of digital libraries. Moreover, automatic keyphrase extraction shows an approach to the summarize of the document. Keyphrase extraction is especially important in the academic publishing, since using it as a technological building block to advise articles to readers, to highlight missing citations or quotations to the authors and also for analyzing research directions. For the time being, keyphrase extraction of documents has become a great request in automatically understanding the topic of academic literature that includes two steps: The keyphrases candidate selection and the keyphrases candidate ranking [Papagiannopoulou, E. and Tsoumakas, G., 2018]. It is easy for human to read and understand the text of a natural language, but it is not practical to process high volume of textual data, group and to regain patterns emotions of the data. Text analysis helps in arranging the data and extracting patterns from textual unstructured data. Also, computerized text analytics approaches are now being carried out in several industries to find out and foretell directions and knowledge patterns in textual data [Chakraborty, G et al., 2014]. There is a vast range of technology focus areas in human language technology include areas such as natural

language processing, text generation and text exploration [Kao, A. and Poteet, S., 2005]. Keywords can be taken into regard as compressed and concentrated text of documents and short forms of their outline; Keyword extraction is a considerable technique for a number of text mining concerning tasks like as document recall web page recapture, document grouping and summarization. The Keyphrases are terms, (single words or phrases) which characterize the important topics of a document [Pianta, E. and Tonelli, S., 2010]. As an outline delegate of the document, key phrases were used widely for gaining essential information. Also, extracting high quality key phrases can be useful for several natural language processing applications, like summarization, information retrieval, question answering, document classification etc. Particularly at the request or subject independent condensation system, it's a necessity module [Bhaskar, P et al., 2012]. Key phrase extraction is divided into manual method and automatic methods. Authors usually specify Key phrases in textbooks. Commonly, most of the manual key phrase extraction is not solid or consistent. In spite of the information and specialists are very skillful, it is hard for a team to classify the content matchmaking and uniquely, and even if by following a standard template [Rheenen, E., 2016]. The best solution for classifying articles by several types and using keyword anchors as hyperlinks between documents makes the user able to fast access concerning material is to list documents related to a primary document's keywords [Jones, S. and Paynter, G.W., 2002]. A topic is a certain subject which we write about it. To identify the topics of the texts, Information Retrieval (IR) the researchers uses the word frequency, sign or hint word, location, and title-keyword techniques. Only word frequency counting is used strongly across various domains among these methods, and the other techniques rely on the unimaginative textual framework or the functional structures of certain domains. Underlying the use of word frequency is the supposition that using a word more times in a text, it becomes more important in that text [Lin, C.Y., 1995].

2. Related Works

In [Kumar, N. and Srinathan, K., 2008], the authors introduced a n-gram based technique to filter key phrases in English documents which was concerning to the scientific domain. The authors computed n-grams (unigram, bigram, and 3-grams) to extract the elected key phrases which are finally assorted basing on the characteristic like term frequency, the position of a key phrase in a document, and position of a key phrase in a sentence. But their system does not use learning or training phase. It does not use the huge domain specific thesaurus in keyphrase extraction and

they need to improve each phase independently to obtain even higher quality results, and develop a mathematical model which can handle the entire language system.

In [Mustafa, A et al., 2009], the authors discussed an implementation or application of information extraction and classification in the text mining application. Author's record that their own application provides exact results only documents which are related to the computer science field and the domain dictionary is loaded with very few words when compared to the amount of words it should be.

In [Sarkar, K et al., 2010], the authors described a key phrase extraction approach using a neural network system to deal with key phrase extraction from scientific articles, and used five features TF*IDF, the position of a phrase that appears first, the length of terms, word length and the links of a phrase to other phrases to identify the key phrases. The authors mentioned that they must develop their system by improving the candidate phrase extraction module of the system.

In [Kaur, J. and Gupta, V., 2010], the authors described the planning that has been presented which may be linked to extract successful key phrases that basically distinguish an article. The authors made a research on different keyword extraction techniques, like the conditional random field algorithms.

In [Hanyurwimfura, D et al., 2012], the authors suggested a text mining approach for extracting the main information from the scientific research paper such as the title, the author names, the main content, and the experimental results that depended on key terms and key sentences to extract the major data which is required by the reader and also storing the separated data in the database for another handling.

In [Menaka, S. and Radha, N., 2013], the authors are proposed a text ranking technique by using keyword extraction and content grouping. In the proposed system, the authors used term frequency, inverse document frequency and WordNet to extract the keywords from documents. The authors suggested a technique using content mining calculations to concentrate on key terms from research papers.

In [Parmar, P. and Ketan, P., 2016], the authors explained that Quality and speed are the key factors that handling an application depends on as a key term extraction algorithm to select the important sentence.

In [Naidu, R et al., 2018], the authors introduced a summarized technique and used it to extract the critical information and supply a summary of the news in the e-Newspaper. In that way they depended on human interference to coach and teach the system to seek out the potential keywords.

The previous research work was mainly interested in key phrase extraction to make summarization, remove essential data and calculate

key terms from the text. Even, as far as we know, there is quite lacking, almost none, in research works that are interested in supplying a computer based assistance to the academic program /course coordinators for properly selection/ clarification of courses' subjects/topics.

3. The Proposed Methodology

The methodology utilized in this paper went through a process of data collection, data processing, and keyphrase extraction. Programming course was selected randomly in two institutions (Institute of Statistical Studies and Research (ISSR), Department of Information Technology, Cairo university and American University of Beirut (AUB), Department of Electrical and Computer Engineering, Beirut, Lebanon. The goal of this method to select topics automatically without the manual work.

RapidMiner tool was used to collect the datasets. RapidMiner is the famous open source software for data mining and supports text mining and other data mining techniques which are used in the combination with text mining. The power and flexibility of RapidMiner is because of the GUI-based IDE (integrated development environment) it is provided for the rapid prototyping, development of data mining models, and it is strong for scripting based on XML (extensible markup language) [Raghuwanshi, R., 2016].

The methodology of the automatic topics selection technique is shown in Fig. 2.

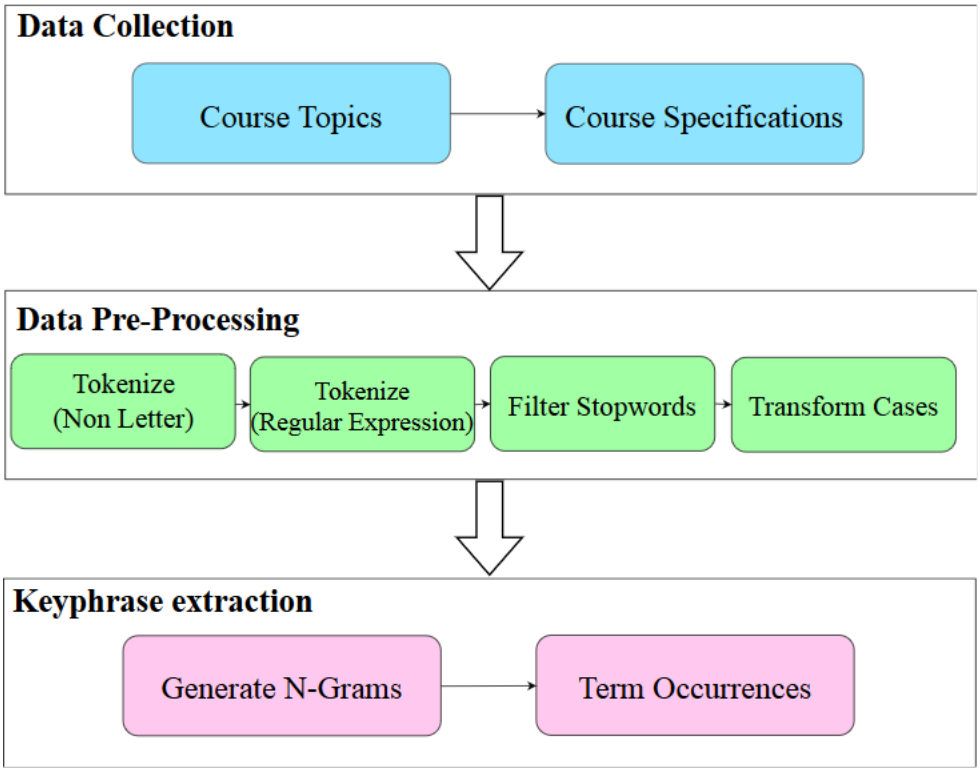


Figure 2: Process Flow Diagram

3.1 Data Collection

In our research this phase contains the following steps:

- **Collecting Course Topics**

The first step is to collect the topics including subtopics of the learning document for the course material that will be educated like textbook, presentation, notes, lecture, ...etc.

- **Collecting Course Specifications**

Course Specifications introduced a detail of what the University agrees to provide in the offering of the course and outlines the expectations on the student, the main information about the course specifications was selected as follows:

- Overall aims of the course
- Catalog description
- Course objectives
- Intended learning outcomes

3.2 Data Pre-Processing

This step makes a clean and divided text into words or tokens. Clean means extracting text containing unnecessary content and removing useless information in the documents, using tokenization, stop word removal and transform cases from data processing.

- **Tokenization**

Tokenization is the process of dismantling the sentences as well as the text file into word delimited by a space or new line. There are two phases of tokenization, the first is to remove the letters using mode (non letters) the second is to remove the regular expression that was found using mode regular expressions ([-!"#$%&'()*+,./:;<=>?@\[\\\]_`{|}~]).

- **Stop Words Removal**

The operator of stop words are the much occurred set of words which does not closely connect information to the text classification task. Then, we have to remove these words from the text documents. The examples of stop words are (a, about, after, an, and, any, are)…etc.

- **Transform Cases**

Its operator is to transform all characters into lowercase letters or uppercase. we use lowercase letter for all data. For example “Exception” transformed to “exception”.

3.3 Keyphrase Extraction

The collection of words remained in the document after all those steps of Pre-Processing are considered as a formal representation of the document, and we call the words in this collection 'terms'. This is our final text body.in this phase we make analyzing for text by collecting the n-grams and select the occurrences of these terms.

- **Generate N-grams**

They are used frequently natural language processing (NLP) and in a task of the text mining . n-grams it is a set of co-occurring words in a specific document. The choice is based on giving the similar frequent two or more words and sentence frequency is counted only for these n-grams. There are 5 types of grams used (unigram, bigram, trigram, four-gram and five-gram).

- **Term Occurrences**

Term Occurrences which uses the presence of a term (value 1) or not (value 0). Furthermore, term occurrences (TO) uses the presence of a term as true or not as false [Ahmad, F. K., 2017]. The frequency of each term in the documents appears as shown in Figure 3 where the occurrence of each term in a two documents is displayed in numeric format. The result is displayed in sorted order by highest of occurrences.

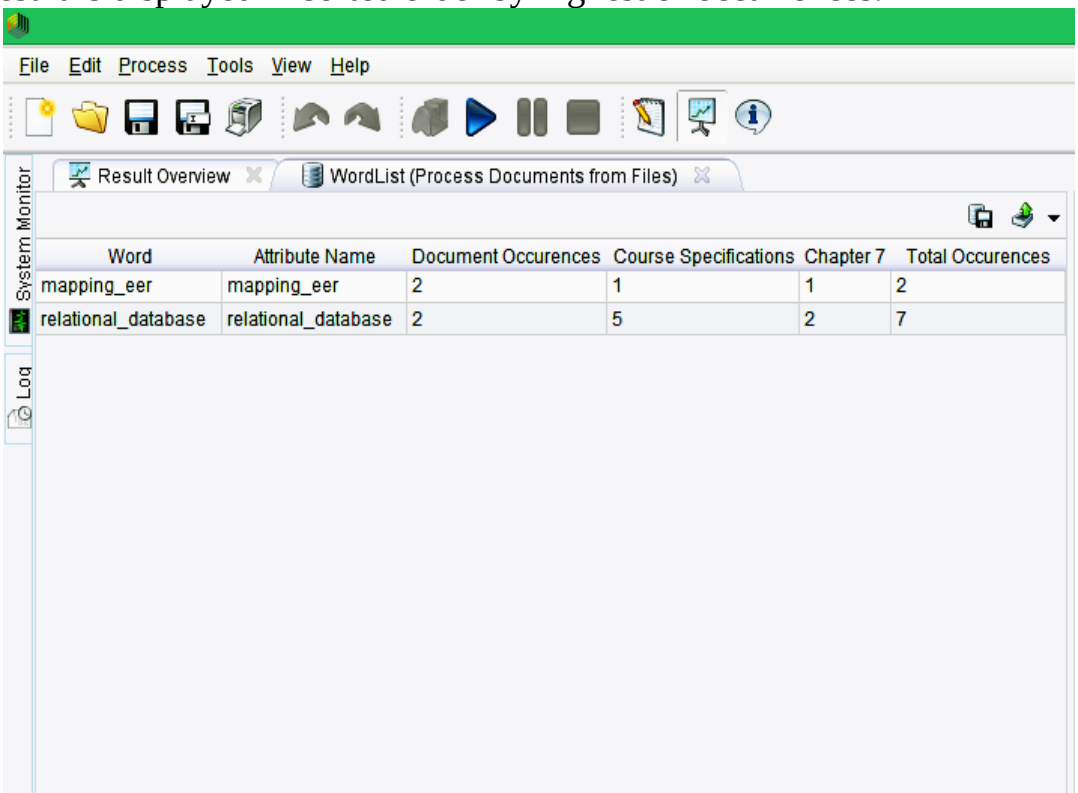


Figure 3: Words and their Frequency Count

During the extraction we compute the frequency of the n-gram types (count of type's occurrences) in our case we ignore words which appear in less than two documents.

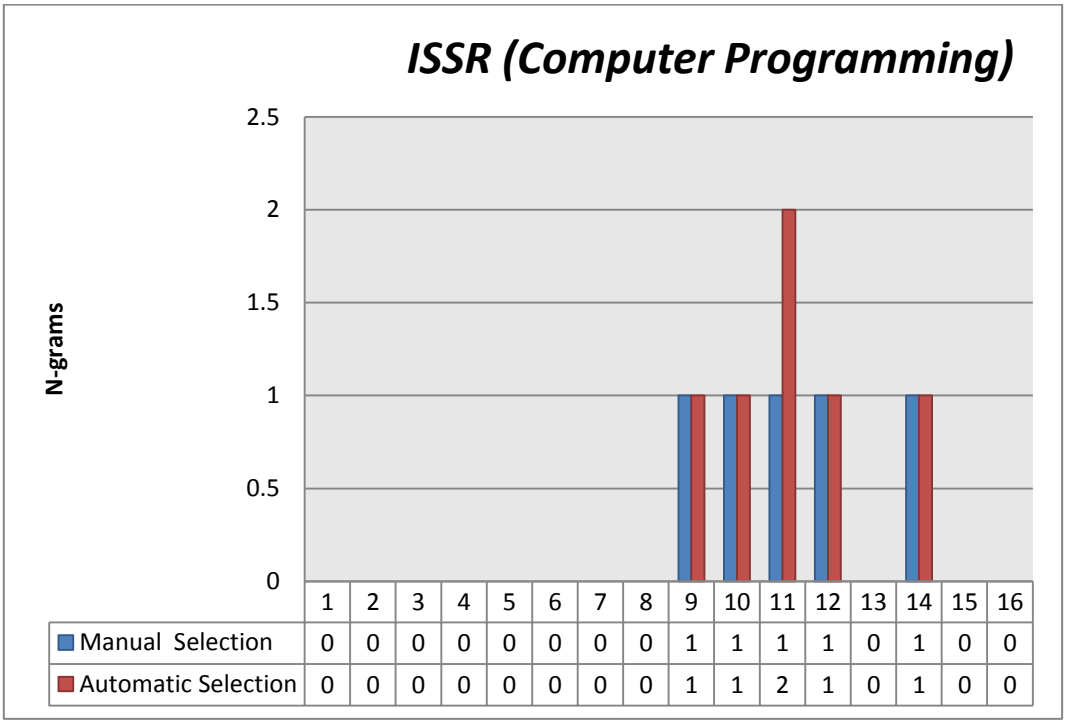
4. Evaluation Results

To evaluate our methodology, we conducted main two types of evaluation tests for Programming course in two different institutions:

- Institute of Statistical Studies and Research (ISSR), department of computer science, Cairo University.
- American University of Beirut (AUB), department of electrical and computer engineering, Beirut, Lebanon.

4.1 Institute of Statistical Studies and Research (ISSR)

In department of computer science in the Institute of Statistical Studies and Research (ISSR) - Cairo University, course code CS523 is taught. The results of applying the automatic topic selection approach for this course show that the five chapters were selected from sixteen chapters. Five chapters among the sixteen chapters are covering nine from ten topics that were found in the course specifications.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Manual Selection	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	0
Automatic Selection	0	0	0	0	0	0	0	0	1	1	2	1	0	1	0	0

Figure 4: Results of Applying the Automatic Topics Selection for Computer Programming Course in the ISSR

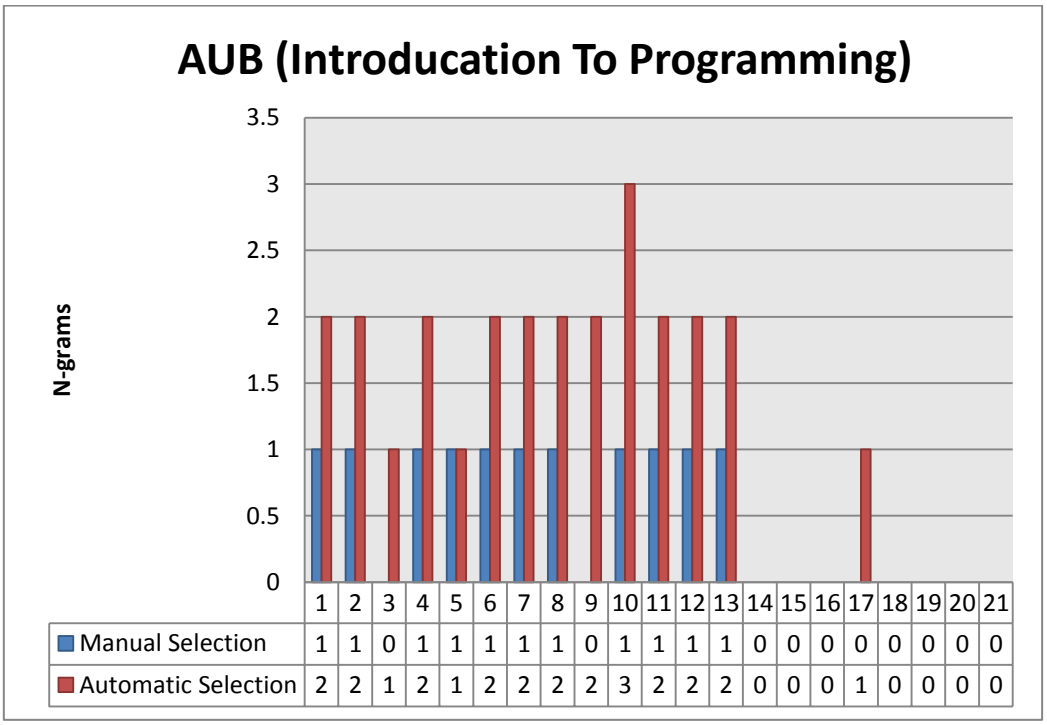
- Comparing with the manual selected topics in the course specifications, topic number ten (Event Handling) not founded in the

textbook. So these topics are required to be mentioned in the other material.

- Based on the obtained results in figure 4 it is clear that: the automatic topic selection approach achieves 90% of manual topics that defined in the course specifications.

4.2 American University of Beirut (AUB)

In the Department of Electrical and Computer Engineering - American University of Beirut (AUB), course code EECE 230 was taught. The results of applying the topic selection approach for this course show that fourteen chapters were selected from twenty one chapters. Fourteen chapters are covering the nine topics founded in the course specifications.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Manual Selection	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0
Automatic Selection	2	2	1	2	1	2	2	2	2	3	2	2	2	0	0	0	1	0	0	0	0

Figure.5 Results of Applying the Automatic Topics Selection Approach for Programming Course in the AUB

- Comparing with the manual topics in the course specifications, we found two topics which were defined in automatic topic selection and were not defined in manual selection. Chapter three achieved the learning outcome number 4 (Know how to input/output data (cin, cout, files…), chapter seventeen achieved the learning outcome number 10 (Know how to implement data abstraction through the use of Classes: public and private members, member functions, constructors, destructors). So these topics required to be selected in the course specifications to achieve course learning outcomes.

- Basing on the obtained results in figure 5 it is clear that: the automatic topic selection approach achieved 85.71% of manual topics that defined in the course specifications.

5. Conclusion

Based on the results that we have acquired, we can say that automatic topics extraction approach can be very powerful when Taking into account the writing of the course specifications accurately to achieve the learning objectives. This approach is able to perform topics extraction in a faster period of time using data mining tool. We believe that it can be a good start for the topic extraction field. With continuous development over time, this study has its edge in processing a number of documents and presenting the topics that achieve the intended learning outcomes for the course coordinators. The course coordinators saved time and effort to select topics for semester courses rather than using the traditional manual method, In addition, they will be sure to apply the full learning outcomes of the courses. In future work we look forward to making our approach give us more accurate results and higher accuracy rate for more topics.

Refrences

Talib, R., Hanif, M.K., Ayesha, S. and Fatima, F.,"Text Mining: Techniques, Applications and Issues." International Journal of Advanced Computer Science & Applications, pp. 414-418, 2016.

Jony, R.I., Rony, R.I., Rahman, M. and Rahat, A., "Big data characteristics, value chain and challenges", In*: Proc. of the Advanced Information and Communication Technology, pp.1-6, 2016.*

Papagiannopoulou, E. and Tsoumakas, G., "Local word vectors guiding keyphrase extraction", Vol.54, No. 6, pp.888-902, 2018.

Chakraborty, G., Pagolu, M. and Garla, S., "Text mining and analysis: practical methods, examples, and case studies using SAS", *SAS Institute Inc*., Cary, North Carolina, USA, 2014.

Kao, A. and Poteet, S., "Text mining and natural language processing: introduction for the special issue", ACM SIGKDD Explorations Newsletter, ACM Press, Vol.7, no. 1, pp.1-2, 2005.

Pianta, E. and Tonelli, S., "A flexible system for keyphrase extraction", In*: Proc. of the 5th international workshop on semantic evaluation, Association for Computational Linguistics*, pp. 170-173, 2010.

Bhaskar, P., Nongmeikapam, K. and Bandyopadhyay, S., "Keyphrase extraction in scientific articles: A supervised approach", In: *Proc. of Computational Linguistic*, pp.17-24, 2012.

Rheenen, E., "Manual versus auto-classification", 2016. *Available:https://www.xillio.com/blog/manual-versus-auto-classification.*

Jones, S. and Paynter, G.W., "Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications", *Journal of the American Society for Information Science and Technology*, Vol.53, no. 8, pp.653-677, 2002.

Lin, C.Y., "Knowledge-based automatic topic identification", *In: Proc. of the 33rd Annual Meeting on Association for Computational Linguistics* (ACL-95), pp. 308-310, 1995.

Kumar, N. and Srinathan, K., "Automatic keyphrase extraction from scientific documents using N-gram filtration technique", *In: Proc. of the eighth ACM symposium on Document engineering*, pp. 199–208, 2008.

Mustafa, A., Akbar, A. and Sultan, A., "Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization", *International Journal of Multimedia and Ubiquitous Engineering*, Vol.4, No.2, pp.183-188, 2009.

Sarkar, K., Nasipuri, M. and Ghose, S., "A New Approach to Keyphrase Extraction Using Neural Networks*", International Journal of Computer Science*, Vol. 7, Issue 2, No. 3, pp.16-25, 2010.

Kaur, J. and Gupta, V., "Effective Approaches For Extraction of Keywords", *IJCSI International Journal of Computer Science*, Volume 7, Issue 6, pp.144-147, 2010.

Hanyurwimfura, D., Liao, B., Njogu, H. and Ndatinya, E., "An automated Cue Word based Text Extraction", *Journal of Convergence Information Technology (JCIT)*, Volume7, Number10, pp. 421-429, 2012.

Menaka, S. and Radha, N., "Text Classification using Keyword Extraction Technique", *International Journal of Advanced Research in Computer Science and Engineering*, Volume 3, Issue 12, pp.734-740, 2013.

Parmar, P. and Ketan, P., "A Survey Paper on Mining Keywords Using Text Summarization Extraction System for Summary Generation over Multiple Documents", *International Journal of Science and Research (IJSR)*, Vol.5, Issue 11, pp.1304-1307, 2016.

Naidu, R., Bharti, S.K., Babu, K.S. and Mohapatra, R.K., "Text summarization with automatic key word extraction in Telugu E-Newspapers", *Smart Computing and Informatics*, *Springer*, *Singapore*, pp.555-564, 2018.

Raghuwanshi, R., "A Comparative Study of Classification Techniques for Fire Data Set." International Journal of Computer Science and Information Technologies, Vol. 7 (1), 2016, pp. 78-82.

Ahmad, F. K., "Comparative Analysis of Feature Extraction Techniques for Event Detection from News Channels' Facebook Page." Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 9, no. 1-2, pp. 13-17, 2017.

An Approach for Discovering Hotspots Using GIS and Fuzzy Set

Sherif M. Akl [1] **Nagy Ramadan** [2] **Hesham A. Hefny** [3]

Abstract

Recently, Detecting hotspot had attracted many researchers. There are some issues related to discovering the level of hotspot, such as dealing with uncertainty since the real world events comes with incomplete data and inexact number. This Paper introduces a new method for the hotspot ranking process for the areas that have a greater than average number of disorder events. Nowadays, decision makers pay more attentions for detecting crime hotspots to allow police personnel to plan effectively for areas which need more concentration, determine mitigation priorities and analyze historical events. This work suggests a fuzzy based method for detecting crimes hot spots. This approach is depending on fuzzy geographic information system (FGIS) model which is formed of three main phases which are: (1) preprocessing, (2) Fuzzification, and (3) GIS visualization. Visualizing the hot locations on map using GIS technology enables the decision makers to build a lot of shapes for criminal analysis and solutions. Also, it determines the most gravity location areas by ranking the locations according to the historical news of committed crimes. The proposed approach tries to help people and decision makers knowing accurately hot places and its ranks.

Keywords*: Hotspot, GIS, Fuzzy sets, Fuzzy membership, FGIS.*

1. Introduction

Crime is a phenomenon which is universal in its varying forms in all cultures and societies. The warning increased in the rate of criminal activities entire the world [Ahmed. M and Salihu. R.S, 2013]. Today's world is suffering from huge rate of crimes, which increased during the last few years. The current studies estimated that the global rate was 7.6 intentional homicides per 100,000 inhabitants for 2004. UNODC (United Nations Office on Drugs and Crime) reported a global average intentional homicide rate of 6.2 per 100,000 populations for 2012, Table 1 shows the rate of crimes round the world [UNODC.ORG].

[1] Information Systems and Technology Dept., Institute of Statistical Studies & Research (ISSR), Cairo University, Egypt.

[2] Information Systems and Technology Dept., Institute of Statistical Studies & Research (ISSR), Cairo University, Egypt.

[3] Computer Science Dept., Institute of Statistical Studies & Research (ISSR), Cairo University, Egypt

TABLE 1. UNODC Crime rates, most recent years

Region	Rate (%)	Count
Americas	16.3	157,000
Africa	12.5	135,000
Europe	3.0	22,000
Asia	2.9	122,000

In Egypt, also, people are suffering from increasing crime rate in the recent years [Numbeo.com]; these crimes were spread via large number of places in Egypt as shown in fig (1).

Fig. 1. Crime rates in Egypt based on Numbeo organization.

One of the challenges that face the police agencies is that the vagueness of determining the most important places to be manipulated in concentrated manner, so, researches pays more attentions on innovative technological tools to confront this weakness to reduce the huge number of crimes. Large collection of data is treated with the help of proposed software tools to reach the optimal using of data.

2. Previous Work

The previous related work mainly identifies crime hotspots based on the locations of high crime density. Nidhi and Amit Kumar (2018), proposed an algorithm that applied fuzzy k-means of the crime data which leads to cluster formation in a better way. Their result shows that the authors reduced the number of clusters from 3 to 2 clusters in their case study, in spite of these significant results, this approach has some disadvantages; there is no information available that includes ranking the crime upon its types and ranking the locations according to its severity.

Alejandra A. Lo'pez-Caloca and Carmen Reyes, (2018), prepared a survey to demonstrate the advantages of utilizing fuzzy methods for spatial analysis and image processing applications. For crime analysis, they were able to identify patterns by looking at the geography of the incidents and identifying hotspots. Zones with a high spatial concentration of schools were also identified, as well as the existence of geographic areas needing this service. The models designed and applied in the work allowed to identify different aspects of spatial patterns, where the main elements of the study were part of the geographical landscape. The question remains. Is there a method to identify crime types per location and label this location according to its severity? Also, is there a method to give the expert a detailed report for spreading of some types of crimes in some areas rather than others? The answer for these questions will be introduced during this work.

The rest of this paper proceeds as follows. Section 3 gives basic concepts. Section 4 introduces a proposed approach. Section 5 discusses a real case study on Cairo crimes data set and shows how the approach implemented on the data set, also, it introduces a visualizing model that helps decision makers finding their target locations and crime they interest. Section 6 summarizes the research as conclusion.

3. Basic Concepts

The presented approach in this paper is fuzzy based approach. Fuzzy set first introduced by Lotfi A. Zadeh [1965]. It differs from the classical notion of set by allowing the gradual assessment of the membership of elements. This is described with the aid of a membership function valued in the real unit interval [0; 1] [Zadeh, 1988].

- **Fuzzy logic** is one of the techniques of soft computing, i.e. computational methods tolerant to suboptimality, impreciseness (vagueness) and partial truth and giving quick, simple and sufficiently good solutions [Stoffel, 2010].
- **Linguistic Variables** are variables that, instead of numerical values, consist of linguistic terms [Svensson, 2016]. Consider the linguistic variable gravity which may consist of the terms high, medium and low. These linguistic terms can each be described using a fuzzy set. Using linguistic variables in this way makes it possible to describe vague and ambiguous concepts in a way that is understandable by machines [Svensson, 2016].

- **Fuzzy System** is one of the most commonly used tools for the data collection and filtering. The filtering mechanism will be accomplished by the use fuzzy rules. These rules are IF-THEN rules. Fuzzy system consists of input stage, processing stage and then output stage. An example for fuzzy rules: as If temperature is "low" THEN heater is "High" [Stoffel, 2010].

  The main idea behind a fuzzy system is to use the concept of linguistic variables to make decisions based on fuzzy rules and thereby get a better response compared to a system using crisp values [Svensson, 2016].

- **GIS** is one of the most influential tools for facilitating and exploration of the spatial distribution of crime. The fundamental strength of GIS over traditional crime analytical tools and methods is the ability to visualize, analyze and explain the criminal activity in a spatial context [Ahmed. M and Salihu. R.S, 2013]. Environmental Systems Research Institute, (ESRI) California (1990), defined GIS as an organized collection of computer hardware, software and personnel to efficiently capture, store, update, manipulate, analyze and display all forms of geographically referenced information [Ahmed. M and Salihu. R.S, 2013]. GIS uses geography and computer-generated maps as an interface for integrating and accessing massive amounts of location-based information.

  GIS is the optimal solution used to answer the question, Where are the highest concentrations of crimes? Several algorithms are available to calculate the areas of highest density in a point distribution. In addition the use of maps by the police using GIS and remotely sensed data allows analysts to identify hot spots, along with other trends and patterns.

4. Proposed Approach

Fuzzy geographical information systems (FGIS) model interfaces GIS with fuzzy models.

FGIS Model as shown in fig (2) is formed of three main phases are: (1) preprocessing, (2) Fuzzification, and (3) GIS. The main three phases described as follows:

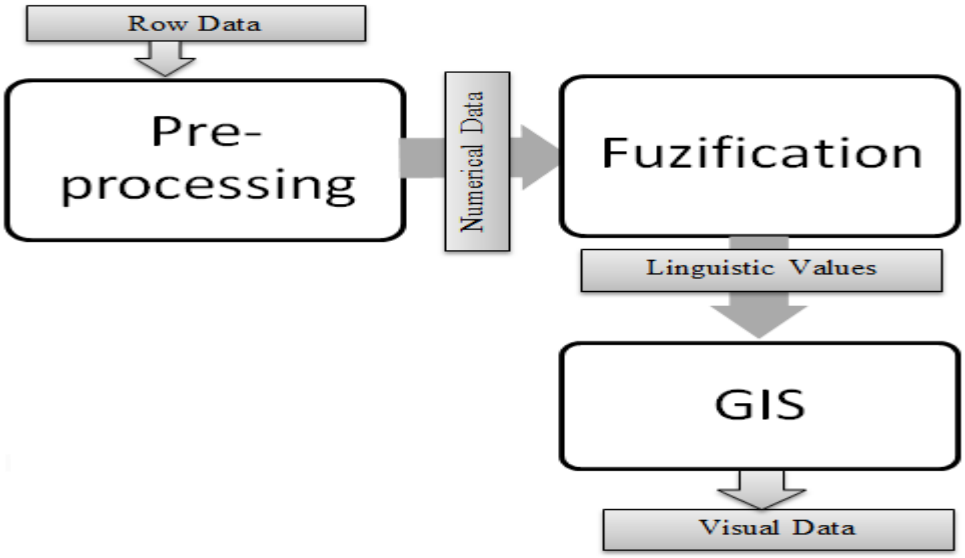


Fig. 2. FGIS Model

4.1 Preprocessing phase: The collected data set represented crimes occurred in 2016 [Memonitor.com], it was comprised of 1000 total instances of Cairo locations. Figures (3) and (4) show samples of collected data. The following mechanisms were used to select the final set of attributes:

- All attributes with the same type collected together such as Violent Crimes, Murders, Robberies, and Assaults.
- Redundancy was removed.
- Categorizing each crime type according to location by calculating the number of each crime in the location as shown in figure (5).

Fig. 3. Sample of raw data.

Fig. 4. Sample of Cairo crime collected data raw of year 2016

Commercial	Human rights	Accidental	Robbery	Drugs	Killing	Terrorism	City
16	28	72	60	6	90	6	Tagamo
15	12	30	210	165	86	3	Polaq
9	18	20	96	28	24	2	Qualiob
6	10	15	60	16	19	1	Qasrelnile
3	8	22	15	9	8	1	Qasr eleiny
10	23	42	32	72	32	7	Rodelfarag
52	38	25	10	37	20	3	awsem
82	66	107	88	91	112	9	Eldarb
259	140	205	210	124	180	29	Nasr city
90	49	58	38	112	70	3	elbasaten
26	14	24	85	48	60	1	elkattamia
104	68	130	220	180	170	35	Old Cairo
45	68	44	58	99	50	6	Abdeen
28	36	65	29	57	48	3	Alsharabia
49	30	45	81	117	21	2	bab alkhalk
46	10	33	49	41	18	7	al mokattam
94	35	58	118	102	44	1	polak ela
83	24	105	201	130	109	22	marj
42	26	75	59	82	28	3	mansheyet r
90	49	48	85	100	80	5	Zawya
25	42	70	56	63	40	2	sakr korish
66	60	102	84	109	104	11	Helwan
195	42	49	189	115	68	2	Mosky
10	6	24	28	40	24	1	Hadaek
56	44	69	130	109	95	12	Ain Shams

Fig. 5. Sample of Cairo crime numerical data after data preprocessing

4.2 Fuzzification phase: Since the numbers of crimes refer to the degree of severity, these numbers will convert to one of the linguistic values that contained into the linguistic variable severity.

Severity = {Very Low, Low, Medium, High, Very High}

In this phase the data is converted from numerical values to fuzzy linguistic values according to the following formulas.

The formula to calculate (Very Low) values will be as following:

$$\mu\, Very\ Low\ (X) = \begin{cases} 0, & min\ +\ 2d\ <=\ X\ <=\ min \\ \dfrac{(\min -\ x)}{d} + 2, & min + d\ <=\ X\ <= min + 2\,d \\ 1 & min\ <=\ X\ <=\ min\ +\ d \end{cases} \quad (1)$$

The formula to calculate (Low) values will be as following:

$$\mu\ Low\ (X) = \begin{cases} 0, & min\ +\ 3d\ <=\ X\ <=\ min\ +\ d \\ \left(\frac{1}{d}\right) * (x\ -\ min) - 1, & min + d\ <=\ X\ <=\ min + 2\ d \\ 1\ +\ \left(\frac{1}{d}\right) * (min - x) & min + 2d\ <=\ X\ <= min + 3\ d \\ 1 & X\ =\ min\ +\ 2\ d \end{cases} \quad (2)$$

The formula to calculate (Medium) values will be as following:

$$\mu\ Medium\ (X) = \begin{cases} 0, & \min\ +\ 4\mathrm{d}\ <=\ X\ <=\ min\ +\ 2\mathrm{d} \\ \left(\frac{1}{\mathrm{d}}\right) * (\mathrm{x}\ -\ \min) + 3, & \min + 2\mathrm{d}\ <=\ X\ <=\ min + 4\ d \\ \left(\frac{1}{\mathrm{d}}\right) * (\min - \mathrm{x}) + 4 & \min + 3\mathrm{d}\ <=\ X\ <= \min + 4\,\mathrm{d} \\ 1 & \mathrm{X}\ =\ \min\ +\ 3\,\mathrm{d} \end{cases} \quad (3)$$

The formula to calculate (High) values will be as following:

$$\mu\ High\ (X) = \begin{cases} 0, & min\ +\ 5d\ <=\ X\ <=\ min\ +\ 3d \\ \left(\frac{1}{d}\right) * (x\ -\ min) + 4, & min + 3d\ <=\ X\ <=\ min + 4\ d \\ -3\ +\ (1/d)\ *\ (x - min) & min + 4d\ <=\ X\ <= min + 5\ d \\ 1 & X\ =\ min\ +\ 4\ d \end{cases} \quad (4)$$

The formula to calculate (Very Low) values will be as following:

$$\mu\ Very\ High\ (X) = \begin{cases} 0, & X\ <=\ min\ +\ 4d \\ \left(\frac{1}{d}\right) * (x\ -\ min) + 5, & min + 4d\ <=\ X\ <=\ min + 5\ d \\ 1 & X >= \ min\ +\ 5d \end{cases} \quad (5)$$

Where:

$$d = (max - min) / 5$$

d: represents the distance between each two boundaries on the membership function.

4.3 GIS interface phase: This research adapted some existed software for this phase; this work depended on Arc map software application which is the product of Esri Company that works in GIS field. The work also used the Arc catalog as a database. The hot spots data was saved in Arc catalog

and used by Arc map to display thematic maps in a format that enables the decision makers building their perception upon results obtained.

Linguistic values are used to describe the degrees of severity of target location , Also, different colors was assigned to each linguistic value to express the degree of severity as show in table (3).

TABLE 2. Linguistic values that describes the degree of severity

Rank	Linguistic values
1	Very Low (VL)
2	Low (L)
3	Medium (M)
4	High (H)
5	Very High (VH)

TABLE 3. Ranked color upon its seriousness

Rank	Description	Color
5	Cities which has Very High Severity crimes	
4	Cities which has High Severity crimes	
3	Cities which has Medium Severity crimes	
2	Cities which has Low Severity crimes	
1	Cities which has Very Low Severity crimes	

5. Experimental results & Discussion

The experimental results will be shown during the next steps:

5.1 Dataset and its Characteristics

The dataset information used in our experimental research is acquired from Middle East Monitor website [Memonitor.com]. Then data were cleansed manually, after that, they were sited and aggregated in Excel file. The data were classified into various types according to multiple classifications such as regions, years, crime types, formats and characteristics. The data of each region were collected separately. Data set includes offenses crime and incidents crime. 98% of the crimes in the dataset occurred in the year 2016.

The dataset is composed data of crimes in 36 Towns with 81 of crime instances. The crime category was ranked by the experts, who ranked crimes from the most seriousness on public security, Such as Terrorism, Murder and Theft, downward.

TABLE 4. Cairo crimes numerical data

Postal code	Locations	Terrorism	Killing	Drugs	Robbery	Accidental Accident	Human rights	Commercial crimes
11835	Tagamoaa	6	90	6	60	72	28	16
11567	Kasrelnil	1	19	16	60	15	10	6
11619	Rodelfarag	7	32	72	32	42	23	10
11639	Al-darbelahmar	9	112	91	88	107	66	82
11816	Nasrcity	29	180	124	210	205	140	259
11634	Al-basateen	3	70	112	38	58	49	90
11632	Masrelqademia	35	170	180	220	130	68	104
11613	Abdeen	6	50	99	58	44	68	45
11653	Al-Sharabia	3	48	57	29	65	36	28
11638	Babelkhalq	2	21	117	81	45	30	49
11571	Elmokattam	7	18	41	49	33	10	46
11611	Bolakaboelela	1	44	102	118	58	35	94
11833	Al-marg	22	109	130	201	105	24	83
11732	Manshyetnasser	3	28	82	59	75	26	42
11659	Al-zawyaelhamra	5	80	100	85	48	49	90
11931	Sakrquorysh	2	40	63	56	70	42	25
11511	Al-musky	2	68	115	189	49	42	195
11663	Hadyekelqobba	1	24	40	28	24	6	10
11782	Ainshams	12	95	109	130	69	44	56
11581	Al-Khalifa	1	86	110	71	28	64	38
11745	Darelsalam	4	114	120	89	82	58	36
11684	Al-dowyqa	1	43	98	86	28	40	11
11522	Ramses	4	28	91	219	108	89	188

5.2 Fuzzification

A Fuzzification process for numerical data that came out from the previous phase is the most important phase in this approach. Fuzzy membership function represents the seriousness of each crime type according to the expert's point of view. The expert's job is evaluating each number of certain crimes by the most suitable linguistic value to describe the severity of each crime type in location.

Fig (6) show linguistic values (VL, L, M, H, and VH) on the membership function.

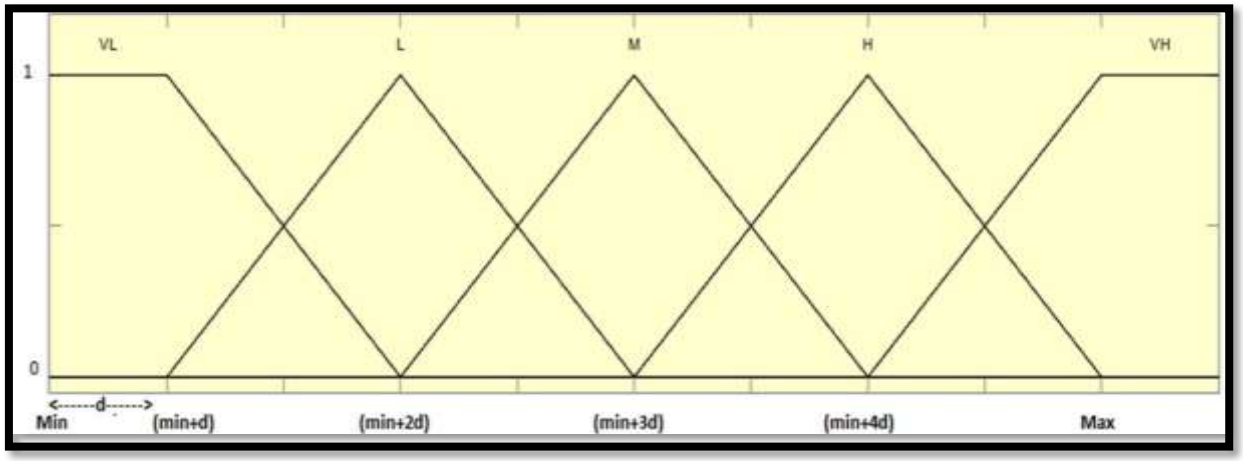


Fig. 6. Fuzzy membership function for Crime severity

5.3 GIS

The results was described in Arc Maps as shown in figures (7) and (8)

The map shows that the locations such as (Nasr City, Shubra, AL musky, etc.) are labeled by red color which means High severity, also shows locations such as are labeled by green color which means that these areas less severity than previous one.

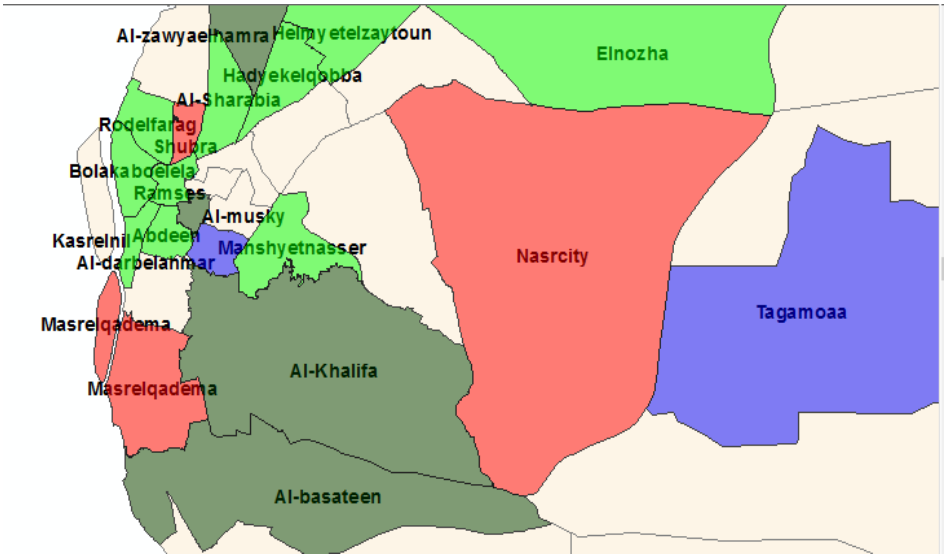


Fig. 7. High Hotspot locations in first quarter of year 2016

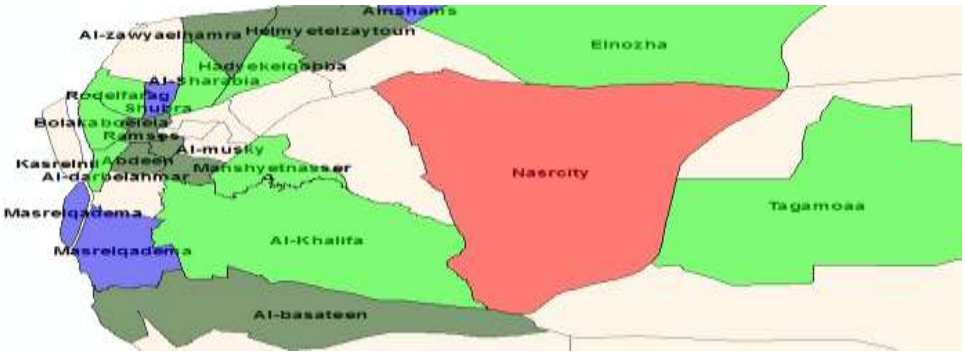


Fig. 8. High Hotspot locations in last quarter of year 2016

6. Conclusion

Detecting hotspot crimes using fuzzy set is capable of classifying crime location based on its severity. Crime mapping and spatial analysis are significant tools for analysis and visualizing crime data. Using GIS gives a good visualization for crimes locations; the suggested approach facilitates spotting high severity locations that needs more attention from those who are interested in fighting against crime. Ranking the crime hotspots indicates the level of seriousness for these locations. The experts and decision makers in a crimes field can easily use this visualized approach and modify the membership functions according to their needs.

References

Ahmed. M and Salihu. R.S, (2013) "Spatiotemporal Pattern of Crime Using Geographic Information System (GIS) Approach in Dala L.G.A of Kano State, Nigeria", American Journal of Engineering Research (AJER), pp. 51–58.

Unodc.org[Online].Available:https://dataunodc.un.org/crime/intentional-homicide-victims.

Numbeo.com[Online].Available:https://www.numbeo.com/crime/country_result.jsp?country=Egypt

Di Martino, Ferdinando, and Sessa. S, (2013) "Hotspots detection in spatial analysis via the extended gustafson-kessel algorithm." Advances in Fuzzy Systems.

Tomar, Nidhi, and Manjhvar. A. K, (2018) "Role of Clustering in Crime Detection: Application of Fuzzy K-means." Advances in Computer and Computational Sciences. Springer, Singapore, 591-599.

Zadeh, L. A, (1988)" Fuzzy Logic", IEEE, pp. 83-93, April.

Stoffel. K, Cotofrei. P and Han. D, (2010), "Fuzzy methods for forensic data analysis." Soft Computing and Pattern Recognition (SoCPaR), International Conference of. IEEE.

Kaur, Navjot, (2016), "Data Mining Techniques used in Crime Analysis:-A Review," International Research Journal of Engineering and Technology (IRJET) Volume 3.

Svensson. S, (2016), "Implementing a Fuzzy Classifier and Improving its Accuracy using Genetic Algorithms"**,** Mälardalen University, M.Sc. Program in Computer, Engineering and Electronics.

Kumar. P. A, and Tiwari. R (2016) "A Review Paper on Geolocation Based Employee Attendance Monitoring System Using Geotagging." provider (ISP) 3: 4G.

Memonitor.com. [Online]. Available: https:// www.memonitor.com. [Last visited: 2- January-2017].

Changjun. F (2014), "A fuzzy clustering algorithm to detect criminals without prior information." Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE Press.

E. POLAT. (2007)"Spatio-temporal crime prediction model based on analysis of crime clusters", Middle East Technical University, graduate school of natural and applied sciences, September.

Hefny, H. A. (2014), "Introduction to Fuzzy Sets and Fuzzy Logic", Lecture Notes.

Fully Automated Detection of Breast Cancer from Digital Mammograms using GMM and SVM

N.R. ElSokkary , A.A. Arafa, A.H. Asad, H.A. Hefny

Abstract:

Computer Aided Detection systems (CAD) are the recent trend in digital mammography which helps radiologist in detection of breast cancer. Recent studies are showing that Segmentation of ROI is an important and critical step and challenging task in the development of CAD system for the detection of breast cancer. The early detection of breast cancer can save many women lives. The digital mammography considered one of the best diagnosis tools in early stages. Therefore, this work proposes a CAD system for breast cancer detection from digital mammography based on Gaussian mixture model (GMM) and support vector machine (SVM). The best contribution of our proposed system is the usage of GMM for the first time in the literature for ROI segmentation. Besides, the classification between three classes of tissue normal, benign and malignant. Texture features and shape features are extracted from the segmented Region of interest and fed into Support Vector Machine (SVM) for classification. The experiments show overall average classification accuracy of 83.3% for detecting normal, malignant and benign on randomly chosen 90 cases from the benchmark mini-MIAS dataset. Additional evaluation metric which is confusion matrix was also used to test and evaluate the proposed CAD system overall balanced performance.

Keywords:

Breast cancer diagnosis,CAD system,GMM, EM algorithm, Mammogram,Mini-MIAS, SVM classifier.

I. Introduction :

Nowadays, breast cancer is one of the most common cancers among women in wide of world. Every year, a large number of women lose their life because of breast cancer. Each year 1.2 million people are diagnosed with breast cancer in the world, with about 400 thousand deaths. There is no known way saving lives from breast cancer. Many researches have shown that early detection could save lives and increase Increased Treatment opportunities. There are several screening tests are available to detect the cancer such as ultrasound, mammography and magnetic resonance imaging (MRI). Mammography is one of the most effective tools for early detection of cancers, which increases the survivability of patients; it is a specific type of the imaging that uses a low dose X-ray system and high resolution film for examination of the breast. And it is based on the difference in absorption of X- rays between the various tissue components of the breast such as fat, tumor tissue, glandular tissue and calcifications M. Guo, M. Dong, Z ., et al.2015.

In order to help radiologists interpret mammograms; many computer-aided detection (CAD) systems have been developed in recent years.The CAD is a strategy

[1] Radiation Engineering Department, National Center for Radiation Research and Technology (NCRRT), Atomic Energy Authority, 3 Ahmed Elzomor St., Nasr City, Cairo, Egypt.
[2]Computer Science Department, Institute of statistical studies and researches (ISSR)-Cairo University

alternative to double reading to help radiologists in the interpretation of mammograms and improve the performance of mammography in detecting breast cancer.

The purpose of CAD systems is to detect suspicious areas within the breast that the radiologist should look at. These systems usually involve detecting candidate lesions: masses or micro calcifications (MCs). The purpose of CAD systems is to help assess the risk of cancer based on the analysis of the detected (and possibly segmented) lesions. These systems usually involve characterizing the suspected lesions and building a classifier based on these characterizations.

The CAD system developed here will detect the mass in mammograms and help radiologists to identify the regions of the problem. Hence CAD system for breast cancer has been become out to be a successful supplemental tool in the combat against breast cancers. It improves the detection rate especially in younger women, in which tumors are hidden because of dense breast tissue R. Ghongade and D. Wakde,2017. Among all the abnormalities present in mammogram, detection and diagnosis of masses as benign and malignant are one of the most challenging tasks, because of their variability in shape, size, and margin, and occlusion within dense breast tissue. Although benign mass is generally characterized by its smooth well-circumscribed boundary and round or oval and obscured shaped low density regions and malignant mass is characterized by its speculated boundary, ill-defined, irregular or stellate shaped and high-density regions, exception occurs very frequently. However many research has had I relied mainly on many features of shape, margin, and texture to classify the masses A. Midya and J. Chakraborty,2015.

In this paper, a proposed method for the detection of breast cancer was developed, based on Gaussian mixture model (GMM) clustering algorithm for region-of-interest (ROI) segmentation and support vector machine (SVM)for the ROI classification on randomly selected 60 mammogram images from the benchmark Mini-MIAS dataset. This proposed system consists of five steps. Firstly; the pre-processing step for removing the background and label from mammogram images. Secondly; the pectoral muscle removal. Third; the segmentation of the ROI from the breast region using GMM. Forth; the Extraction of the features set for the segmented ROI. Finally; the classification of the ROI into normal or abnormal using binary SVM.

The paper is organized as follows: Section II present a literature review for previous works for segmentation and classification for breast cancer. Section III contains general overview to the proposed of our CAD system and the materials using and the system. Section IV contains the Experiment, in section V results and discussion. Section VI presents conclusions Finally, in section VII suggestions for future work.

II. LITERATURE SURVEY :

In the literature, many mass detection and segmentation methods have been developed for the detection of regions of interest (ROI) in mammograms. Soulami, K.B., M.N. Saidi, and A. Tamtaoui, 2017. Used the metaheuristic algorithm particle swarm optimization (PSO) to extract the ROI and then using Fourier transform and Gray Level Co-Occurrence Matrix (GLCM) to extract shape and texture features from the abnormalities. The classification of the detected abnormalities is carried out through the SVM, which classifies the segmented region into normal and abnormal. Their algorithm tested on mini-MIAS dataset and achieved 83.87% accuracy. from Weaknesses for this system the method Which are applied for remove pectoral muscle area from the mammogram images still have some drawbacks when it comes

to the removal of pectoral muscle in dense mammograms , and the level of Entropy thresholding is chosen manually To avoid the over segmentation of the breast. Elmoufidi, A., et al.2014. used dynamic K-means clustering algorithm with local Binary pattern (LBP) for the detection of ROI in mammograms. In this system the contrast enhancement step is necessary; therefore a two-dimensional (2D) median filtering was used for step of enhancement of contrast for mammogram images. Their algorithm tested on mammograms from the MIAS dataset and the results proved that the algorithm is efficient for detection of ROIs in mammography with a mean of 85% and 2.84% false positives per image (FPPI).from the Weaknesses in this work it is the Dependence on the applied the step of enhancement for increase the contrast of images and improve the image for preparation for the following steps in the system.

Guo, M., et al.2015. proposed a new snake model for ROI segmentation to solve the classical vector field convolution (VFC) problem in image segmentation. Combined with Laplacian operator morphological filter, a new external force is calculated by convolving the edge map with the user-defined vector field kernel. They used MIAS database for testing their method and compared it with gradient vector flow (GVF) snake and classical VFC snake. Their proposed method shows its advantages, including the reduced computational cost and better performance. Disadvantages in this work it is depend on find more accurate boundaries of the tumor regions only for the abnormal cases. and the evaluation for their method based on only compare contour iteration time and computational cost with other methods . Ismahan, H., F. Amel, and B. Abdelhafid,2015. presented an effective approach for ROI segmentation based on mathematical morphology for detection of masses in digitized mammograms. Multiple morphological watersheds are performed for detection and localization of various masses types in mammograms to avoid the problem of over-segmentation resulted from single morphological watershed. The algorithms have been evaluated on a set of 38 mammograms from MIAS dataset. In addition, it is compared to the manual detection marked by radiologists. The obtained results show promising performances of their proposed algorithm. The weakness in this work was the evaluation for system performance which was based on just comparing their work with to the manual detection marked by a radiologist. In addition to that in this work has been relied on enhancement process to improve appearance and increase the contrast of images before segmentation step, and applied two techniques in segmentation step for improve segmentation result.

Jothilakshmi, G. R., and ArunRaaza, 2017. proposed a method for the detection and classification of mass abnormalities in digital mammogram images using multi- SVM classifier. They are using region based segmentation for extracting the ROI from mammogram images. And then, they extract the texture features from the ROI samples using Gray Level-Co-Occurrence Matrices (GLCMs).For classification between malignant and benign samples, an optimum subset of texture features is classified using a multi-SVM classifier. For examination their proposed system they are using a set of 50 mammogram images are acquired from Mini-MIAS database and classifies the images features with the significant dataset. It trains the feature extracted with respect to the significant dataset. The input images contain 25 benign images and 25 malignant images. In which the proposed algorithm detects 23 images as true benign and 2 as false malignant images, also in 25 malignant images the algorithm detects 24 true malignant images and 1 false benign image. The accuracy for the classification technique is estimated to 94%.But in this work It is focus only with the classification between benign tumor and malignant, Algorithm efficiency

was tested on mammogram images that the previous knowledge that images it is abnormal type of tissue. Makandar, Aziz, and Bhagirathi Halalli, 2015. Are using multiple combined techniques for segmentation to segment the suspicious mass from The mammography. The segmentation combined techniques consist of watershed, morphological operations and active contour based segmentation techniques. The aim from their segmentation technique which was proposed to helps minimizing the over segmentation of conventional watershed segmentation technique. Algorithm efficiency was tested on mammogram images from mini-MIAS database. The efficiency of the algorithm was measured by using the efficiency equation for calculating the number of true identification images from total number of abnormal images, and error equation for calculate number of false identification images from total number of abnormal images. The efficiency of the algorithm was reported 95.83% and error is 4.17%.The weakness in this work was the segmentation step which was implemented using more than one technique and it has been implemented on more than one stage aiming to improve the ROI segmentation accuracy.

III. PROPOSED SYSTEM :

The proposed CAD system for breast cancer detection consists of five consecutive steps without any previous knowledge of the breast cancer case no any pre-assistance from radiologist, as shown in figure 1.

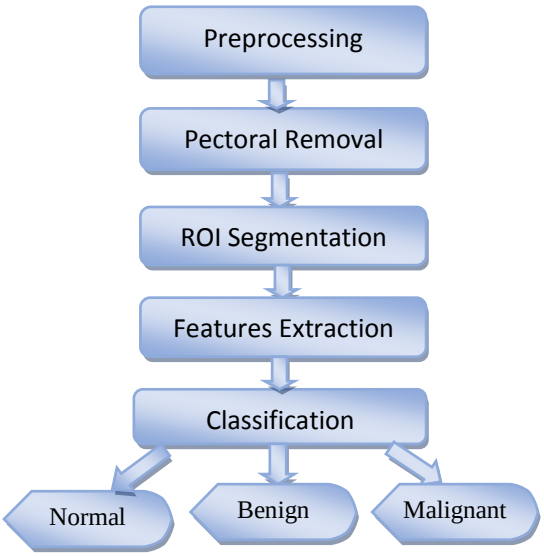


Fig. 1: The proposed CAD system for breast cancer detection

A. Image Preprocessing

Preprocessing is a necessary step for the mammogram images to remove noise. In fact, the breast in mammography image covers only 30% of the entire image. In addition, scanning mammograms generate artifacts as in the case of the Mini-MIAS dataset images (high intensity labels, low intensity labels, scanning artifacts, and artifacts) as shown in figure 2. These artifacts are one of the problems that facing the ROI segmentation process causing many false positives. Therefore in this step, we select automatically only the area containing the breast region and removing these artifacts M. Abdel-Nasser., et al.2015.

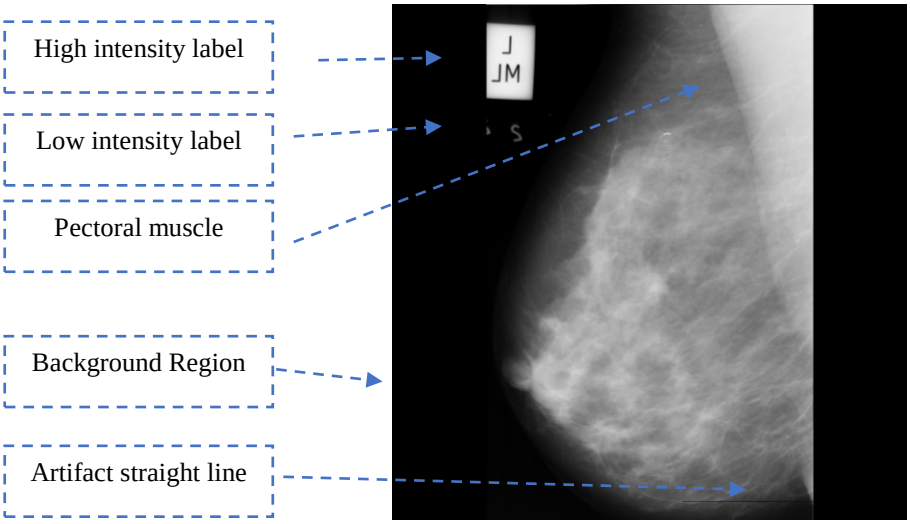


Fig. 2: The elements that constitute a mammogram image

B. Pectoral Removal

In some cases, pectoral muscles are easy to segment from the breast tissue. But in the most cases, the entire pectoral muscle or parts of pectoral muscle are not visible and not easy to be removed**.** In this section, a fully automatic pectoral muscle removal is proposed for successful identification of pectoral region.

Muscle removal method is based on the following properties: firstly, the pectoral muscle is appears in the upper right or left corner of the mammogram. Secondly, the pectoral muscle appears a triangular shaped region. Thirdly, the pectoral muscle area decreasing in size from top to bottom. Finally, the pectoral muscle represents the more density region in the mammogram and it appears brighter than breast tissue. These properties influence the detection of the suspicious ROI area due to its features similarity to the abnormalities and hence need to be removed. For this purpose we proposed a new simple method for automatically removing the pectoral muscle region using the following steps:

1. Change image orientation: The orientation of all images is checked and changed (if necessary) to assure that the chest muscle is located at the top left corner of the image.
2. Filtering the image rows from top and left. This step searches the rows for the valid values that assure the start of pectoral region. The image is then shifted to the top for discarding the existence of bad rows. The same process is done for the left column of the image. Hence, the first point of the pectoral region is asserted to be at (0,0). That point represents the right-angled vertex of the pectoral muscle triangle region.
3. Hence, the first point of the pectoral region is asserted to be at (0,0). That point represents the right-angled vertex of the pectoral muscle triangle region.
4. Calculate second point in the triangle by finding first point exceeds the mean value of the entire row.

5. Calculate third point in the triangle by finding first point exceeds the mean value of the entire column.
6. Apply mask for pectoral muscle region by using the three points which have been selected from previous steps.
7. Subtract the pectoral muscle region from the original image.
8. As shown in figure 3, the pectoral muscle region was accurately removed from the original image after changing the image orientation.

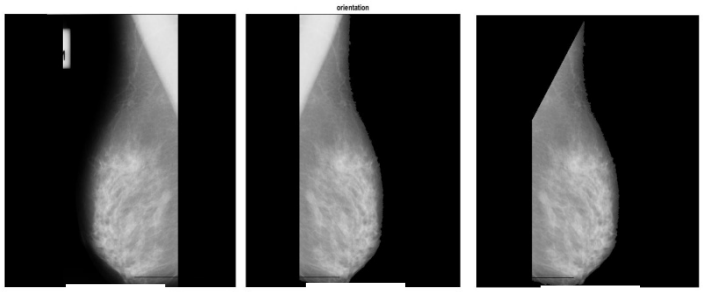


Fig. 3: (a) The original image "mdb081'. (b) Afterpreprocessing and image orientation change.(c)After pectoral muscle region removal.

C. Segmentation

The segmentation step is a critical and very important process because of all subsequent steps (features extraction and classification) is dependent on its result.However, GMM is used to extract the ROI from the breast tissue; the GMM parameters were estimated using the expectation maximization (EM) algorithm M. Guo, M. Dong, Z ., et al.2015. The segmentation of the ROI region is carried out by maximizing the ROI component likelihood. We chose GMM to perform this task because: (1) it is well studied statistical inference techniques. (2) It gives more flexibility in choosing the component distribution. (3) It gives density estimation for each cluster. (4)GMM is able to build soft clustering boundaries, i.e., points in space can belong to any class with a given probability, unlike K-means for example. The GMM is represented by the following equation:

$$f(x_s) = \sum_{i=1}^{k} p_i N(x_i|\mu_i, \sigma_i) \qquad (1)$$

$$(x_i|\mu_i, \sigma_i) = \frac{1}{\sigma i\sqrt{2\pi}} exp\left[-\frac{1}{2\sigma_i^2}(x_s{-}\mu_i)2\right] \qquad (2)$$

The parameters μ_i, σ_i^2, p_i are respectively the mean, variance and the chance that a voxel x_s belongs to the classc_i. There are $c_i, i = 1,2,\dots,k$ classes, where k is the number of classes which is known a priori. This means that$p_i = (x_s \in c_i), 0 < p_i < 1, \sum_i^k = 1, p_i = 1$. Wecan then define the GMM in the following manner:

$$\theta = (p_1, \dots p_k\ , \mu_1, \dots \mu_k, \sigma_1^2, \sigma_2^2) \qquad (3)$$

These parameters must be estimated. One of the most used methods for the estimation of these parameters is the EM algorithm R. Ghongade and D. Wakde,2017.This algorithm is a general iterative technique for computing maximum-likelihood when the observed data can be regarded as incomplete. Each iteration of the EM algorithm consists of two processes: The (E-step) and the (M-step). In the expectation, or E-step, the missing data are estimated given the observed data and current estimate of the model parameters. This is achieved using the conditional expectation, explaining the choice of terminology. In the M-step, the likelihood function is maximized under the assumption that In the M-step, the likelihood function is maximized under the assumption that In the M-step, the likelihood function is maximized under the assumption that the missing data are known. The estimate of the missing data from the E-step is used in allowance of the actual missing data. Convergence is assured since the algorithm is guaranteed to maximize the likelihood at each iteration A. M. Santos., et al.2014., M. A. Mahjoub,2012.

In our proposed GMM based segmentation, the numbers of clusters ***k*** is decided automatically using thresholding to be suitable for each image according to its density which is calculated. After computation of the GMM parameters using the EM algorithm, the mammogram image is segmented into ***k*** clusters regions where each pixel belongs to a cluster. Figure 4 shows the ROIs segmented by GMM from three classes of normal, benign and malignant tissues.

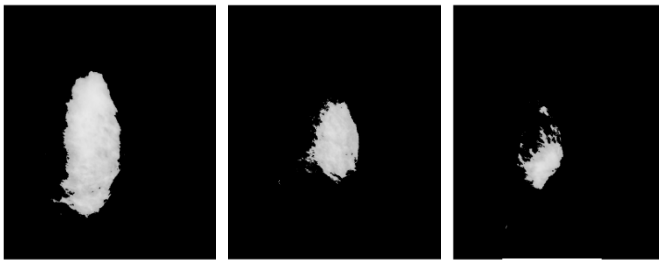


Fig. 4: (a) Normal ROIsegmented by GMM from original mammogram image'mdb003'. (b) BenignROIsegmented by GMM from original image 'mdb212'. (c) Malignant ROI segmented by GMM from original image 'mdb072'.

D. Features Extraction

After the ROI segmentation from breast region, features are extracted to distinguish among the characteristics of them. Generally, the patterns present in a ROI tend to be coarse and low or high in contrast. So the analysis of intensity distribution of a ROI is important for effective classification S. J. S. Gardezi., et al.2017. The pattern present in the ROI exhibits significant texture characteristics, so in this paper fourteen features based on texture, shape and statistics are extracted from segmented ROIs of different sizes in all images. Texture based features are extracted from the ROI using Gray Level Co-Occurrence Matrix (GLCM) A. M. Anter., et 2016. For shape based features, they are four calculated from the ROI; circularity, brightness, compactness,

and volume. Statistics based features are the mean, standard deviation, skewness, kurtosis, smoothness, histogram, correlation, energy and entropy of the ROI M. Abdel-Nasser., et al.2015.

E.Classification

The classification of the ROI as either normal, benign tumor or a malignant one is achieved using the techniques of machine learning. is based on the features that are extracted from the segmented ROI in previous step. The more related and indicative the features are of the ROI in the image, the more accurate the classifier model can perform. Many studies have been conducted on the use of variable machine learning models to classify breast cancer images. A SVM is a supervised learning classifier that discriminates between different classes by finding a hyperplane that separates them. Given a labeled training set of vectors in the form$(x_i; y_i)$; i = 1, 2, … k, where $x_i \in R^n$ are the feature values, $y_i \in \{1, -1\}$is their categories label, n is the number of features (i.e. dimension of the classification problem) and k is the number of samples M. Abdel-Nasser, ., et al.2015., T. Subashini,., et al.2010. The equation of a hyperplane is given by

$$f(x) = x'\beta + b = 0 \qquad (4)$$

Where $\beta \in R^d$ and b is a real number.Finding β and b that minimize ||β|| such that formulating the best separating hyperplane (i.e., the decision boundary). for all test feature points$(x_j; y_j)$,$y_j f(x_j) \geq 1$. On the other hand, the vectors (x_s) that are on the boundary, and satisfy the condition $y_s f(x_s) = 1$are called the support vectors.

In this work, a nonlinear binary SVM classifier is used to classify the segmented ROI as normal, benign and malignant cases of mammogram images based on extracted features as inputs to the classifier.

IV. Experiment :

For the testing and evaluation of the proposed system, we used a publicly available database of the mini-Mammographic Image Analysis Society (MIAS) dataset. This is a publicly available digitized dataset with a subset of 322mammograms. Every image is 1024×1024 pixels gray level image. The tumors have different shapes and sizes. The tumors have been labeled by black rectangle. In general, the contrast of masses to the background is poor. Shapes and sizes of the masses are often irregular, and the borders are ill-defined. Therefore, it is difficult to discriminate between normal and abnormal breast tissues J. Suckling., et al.1994.

Table1: Comparison with recent Malignant-Benign classification methods

Method	Segmentation	Features	classification method	Number of image	Accuracy %
Jothilakshmi, G. R., and ArunRaaz	split and merge technique	Texture features	Multi-SupportVector Machine	50	94%
Ismahan, H., F. Amel, and B. Abdelhafid	watershed transform,morphological gradient		Visual Comparison between region detected and region selected by radiologist.	38	90%
Proposed Method	GMM	Texture, Statistical and shape features	Nonlinear binary SVM	90	83.3%

In this work, a set of ninety mammogram images are randomly chosen from mini-MIAS dataset. into consideration that those samples include images from three different types of breast tissue which are glandular, fatty and dense; for each category of images in Mini-MIAS database; normal, benign and malignant. The images are segmented using GMM with EM algorithm for estimating the parameters of the model. Fourteen features based on texture, shape and statistics are extracted for each segmented ROI .

A SVM classifier is used for classifying segmented ROIs in mammogram images in to the three classes. From the total of 90 mammogram images that are randomly chosen from mini-MIAS dataset, 30 are normal cases, 30 are benign cases and the remaining 30 are malignant cases. The set of 30 cases in each class was randomly divided into two subsets; one set of 10 samples representing 33% was chosen as a training data and the other set of 20 samples representing 67% was used as testing data.

V. Result and Discussion:

Figure 5 presents the results from the first 3 steps of the proposed algorithm as discussed in section III. Each row in that figure shows an example of each studied cases (i.e. normal glandular tissue, benign dense tissue and malignant fatty tissue respectively from top to down). On the other hand, each column of that figure illustrates the output of the corresponding step. Column (a) figure out the mammogram image after the preprocessing step in which the unwanted parts are removed and orientation is fixed. Column (b) illustrates the output of pectoral removal step. Finally, column (c) shows the segmented ROI using GMM.

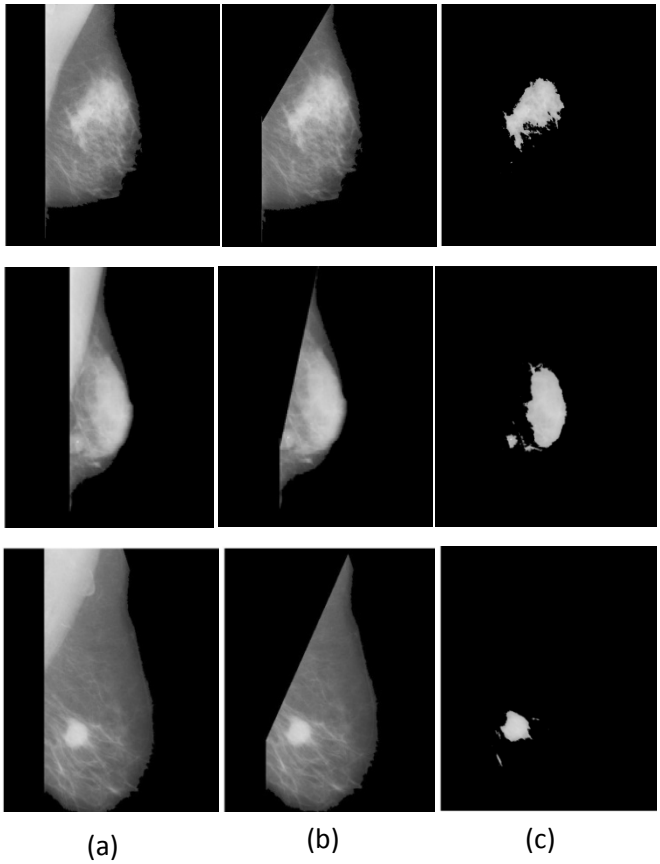


Fig. 5: Result of each stage for three classes. The row (N) is for Normal image'mdb209', row (B) is for Benign image 'mdb290', and row (M) is for Malignant image'mdb028'. The column(a) shows the mammogram image after preprocessing step, column (b) shows the image after pectoral removal step, and column (c) shows the image after ROI segmentation using GMM.

To validate the proposed CAD system ninety mammogram cases were used. Firstly, the extracted fourteen features for all thirty training cases are used to train the classifier. Then, the rest of sixty cases are tested by the classifier. The proposed algorithm detects accurately nineteen cases as true normal with accuracy of 95%, seventeen samples as true benign with accuracy of 85% and fourteen samples as true malignant with accuracy of 70%. Hence, the overall average accuracy was approximately 83.3%.

To the best of our knowledge, the past work was interested only on the classification rate of malignant and benign tumors, in which there is a previous knowledge of abnormality cases .Table 1 shows the comparison of tumors detection with previous work in [Ismahan, H.,et al.2015, Jothilakshmi., et al.2017]. Jothilakshmi., et al.2017 select 50 mammogram images that previously known as abnormal and detect the tumor type at a rate of 94%. Ismahan, H.,et al.2015 using two techniques for segmentation and test type of abnormality at the rate of 90%.

The proposed CAD system has the advantages over the methods in [Ismahan, H.,et

al.2015, Jothilakshmi., et al.2017] and other, that it is fully automated in all its stages and uses less complex and fast computation procedure without any radiologist contribution.

VI. CONCLUSION:

In our proposed system, for the mammogram input image, the segmentation of breast tissue process is implemented in one step by just using GMM for the first time in the literature. The nonlinear SVM is used for classification of the tumor ROI into three classes; normal, benign and malignant tissue. Although the proposed system achieved an overall average accuracy of approximately 67%, but we think it is a promising result taking into consideration that our proposed system was evaluated on just only 60 images which are randomly chosen from mini-MIAS dataset.

VII. FUTURE WORK :

For future work, we plan to operate and evaluate our proposed system on the whole Mini-MIAS dataset. Another useful approach would be replacing the currently used step of pectoral muscle region removal step with GMM for segmenting also the pectoral muscle region from breast boundary, which in turn will give more accurate diagnosis and better performance for the whole system. Applied GMM for breast tissue segmentation and classification of breast tissue density in digital mammograms into three categories Glandular, density and fatty.

References

M. Guo, M. Dong, Z. Wang, Y. Ma, and Y. n. Guo, "A new method for mammographic mass segmentation based on parametric active contour model," in *Wavelet Analysis and Pattern Recognition (ICWAPR), 2015 International Conference on*, 2015, pp. 27-33.

R. Ghongade and D. Wakde, "Detection and classification of breast cancer from digital mammograms using RF and RF-ELM algorithm," in *Electronics, Materials Engineering and Nano-Technology (IEMENTech), 2017 1st International Conference on*, 2017, pp. 1-6.

A. Midya and J. Chakraborty, "Classification of benign and malignant masses in mammograms using multi-resolution analysis of oriented patterns," in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, 2015, pp. 411-414.

K. B. Soulami, M. N. Saidi, and A. Tamtaoui, "A CAD System for the Detection of Abnormalities in the Mammograms Using the Metaheuristic Algorithm Particle Swarm Optimization (PSO)," in *Advances in Ubiquitous Networking 2*, ed: Springer, 2017, pp. 505-517.

A. Elmoufidi, K. El Fahssi, S. Jai-Andaloussi, and A. Sekkaki, "Detection of regions of interest in mammograms by using local binary pattern and dynamic K-means algorithm," *International Journal of Image and Video Processing: Theory and Application,* vol. 1, pp. 2336-0992, 2014.

H. Ismahan, F. Amel, and B. Abdelhafid, "Mass segmentation in mammograms for computer-aided diagnosis of breast cancer," in *Control, Engineering & Information Technology (CEIT), 2015 3rd International Conference on*, 2015, pp. 1-5.

G. Jothilakshmi and A. Raaza, "Effective detection of mass abnormalities and its classification using multi-SVM classifier with digital mammogram images," in *Computer, Communication and Signal Processing (ICCCSP), 2017 International Conference on*, 2017, pp. 1-6.

A. Makandar and B. Halalli, "Combined segmentation technique for suspicious mass detection in mammography," in *Trends in Automation, Communications and Computing Technology (I-TACT-15), 2015 International Conference on*, 2015, pp. 1-5.

M. Abdel-Nasser, H. A. Rashwan, D. Puig, and A. Moreno, "Analysis of tissue abnormality and breast density in mammographic images using a uniform local directional pattern," *Expert Systems with Applications*, vol. 42, pp. 9499-9511, 2015.

R. Nithya and B. Santhi, "Computer-aided diagnosis system for mammogram density measure and classification," *Biomedical Research*, vol. 28, 2017.

A. M. Santos, A. O. de Carvalho Filho, A. C. Silva, A. C. de Paiva, R. A. Nunes, and M. Gattass, "Automatic detection of small lung nodules in 3D CT data using Gaussian mixture models, Tsallis entropy and SVM," *Engineering applications of artificial intelligence*, vol. 36, pp. 27-39, 2014.

M. A. Mahjoub, "Image segmentation by adaptive distance based on EM algorithm," *arXiv preprint arXiv:1204.1629*, 2012.

S. J. S. Gardezi, I. Faye, F. Adjed, N. Kamel, and M. Hussain, "Mammogram Classification Using Chi-Square Distribution on Local Binary Pattern Features," *Journal of Medical Imaging and Health Informatics*, vol. 7, pp. 30-34, 2017.

A. M. Anter and A. E. Hassenian, "Computer aided diagnosis system for mammogram abnormality," in *Medical Imaging in Clinical Applications*, ed: Springer, 2016, pp. 175-191.

T. Subashini, V. Ramalingam, and S. Palanivel, "Automated assessment of breast tissue density in digital mammograms," *Computer Vision and Image Understanding*, vol. 114, pp. 33-43, 2010.

J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, *et al.*, "The mammographic image analysis society digital mammogram database," in *Exerpta Medica. International Congress Series*, 1994, pp. 375-378.

An Intelligent Approach for Predicting the Failure of Agile Software Projects

[1]Ahmed Abdelaziz, [2]Nagy Ramadan Darwish, [3]Hesham A. Hefny, [4]Sherif A. Mazen

Abstract

Predicting the failure of agile software projects is a great challenge faced by software companies. This paper focuses on the using of intelligent techniques such as fuzzy logic, multiple linear regressions to address this challenge. This paper also presents a review of some works related to this area of interest. In this paper, the researchers propose an approach for predicting the failure of agile software projects based on two intelligent techniques: fuzzy logic and multiple linear regressions (MLR). MLR is used to determine crucial failure factors of agile software projects. Fuzzy logic is used for predicting failure of agile software projects.

Keywords-Agile Projects, Intelligent Techniques, Fuzzy Logic, Multiple Linear Regressions.

I. INTRODUCTION

According to the Agile Manifesto (2011), Agile is a framework that concentrates on client value, refined delivery, small teams, self-enterprise and continuous enhancements. Many organizations use agile methodology in software projects to solve traditional methodology problems that lead to lost time, cost and effort [V. Lalsing and et al,2012, J. Li and et al,2012]. Agile is a set of software development operations that are refined, incremental, self-enterprise [John Hunt,2006]. The term Agile was appeared in February 2001 where the Agile Manifesto is formulated. Agile development is a framework for managing the development of software projects, which are based on obtaining early customer feedback with a large number of frequent releases. In Agile development, the work is executed in small time-boxed iterations. Each iteration is viewed as full software development cycle (i.e. planning, requirements analysis, design, coding, unit testing, and acceptance testing). This helps in reducing the overall risk and allows the project to quickly adapt to changes [Ann L. Fruhling and Alvin E. Tarrell, 2008, Laurie Williams and et al,2000, Mike Cohn,2010].

[1]Information Systems and Technology Department, Institute of Statistical Studies and Research, Cairo University, Egypt. ahmed.aziz.1157@gmail.com
[2]Information Systems and Technology Department, Institute of Statistical Studies and Research, Cairo University, Egypt. drnagyd@yahoo.com
[3]Computer and Information Sciences Department, Institute of Statistical Studies and Research, Cairo University, Egypt. Hehefny@ieee.org
[4]Department of Information Systems, Faculty of Computers and Information, Cairo University, Egypt. SMazen@fci-cu.edu.eg

The most commonly used agile methodologies include feature driven development (FDD), Extreme programming (EP), rational unified process (RUP), Adaptive software development (ASD) and scrum [K. N. Rao and et al ,2011, N. R. Darwish and et al,2014, Ioannis G. Stamelos and P. Sfetsos,2007, V. E. Jyothi and K. N. Rao,2011]. In agile projects, the team is usually cross-functional and self-organizing. Most of the agile methodologies have a daily face-to-face communication between team members. The agile lifecycle focuses on test process, customer satisfaction and timely delivery. The agile software development lifecycle is controlled by the incremental process.

Software companies are facing a big challenge for early predicting failure of agile software projects. Most of the studies used traditional methods to predict failure of agile software projects that lead to poor results. There is a need for new methods to predict failure of agile software projects to get more accurate results. This will also help software companies to save cost and time. Furthermore, the factors that affect the success or failure of software projects need to be identified.

The failure of agile software project has many reasons such as weak experience with agile method, insufficient connection between teams and clients, lack of modification, fiscal management, fineness management, inactive sharing of execute management, lack of competent team members and using weak tools [ELZAMLY and B. HUSSIN,2014].

According to CHAOS report, there are three types of software projects as follow:

- Big size software projects
- Moderate size software projects
- Little size software projects

CHAOS report concentrates on three main axes in software projects, as follows:

- Software projects failure
- The major features that cause software projects to fail
- The key ingredients that can minimize project failures

The Standish Group's CHAOS report shows a comparison between agile and traditional waterfall projects. According to this report, agile approaches have more successful projects and less outright failures for every project size. Table 1 shows the results of the most recent report [ELZAMLY and B. HUSSIN,2014].

Table1. Standish Group's CHAOS Report

Size	Method	Successful	Challenged	Failed
All Size Projects	Agile	39%	52%	9%
	Waterfall	11%	60%	20%
Large Size Projects	Agile	18%	59%	23%
	Waterfall	3%	55%	42%
Medium Size Projects	Agile	27%	62%	11%
	Waterfall	7%	68%	25%
Small Size Projects	Agile	58%	38%	4%
	Waterfall	44%	45%	11%

This paper proposes an approach to predict failure of agile software projects. For this purpose, this paper tries to utilize the advantages of intelligent techniques to address this problem. Linear regression analysis is used to identify crucial failure factors. Fuzzy logic is used for predicting the failure of agile software projects.

The rest of the paper is organized as follows: section two presents a background overview; section three introduces the previous work in this research filed; section four introduces the proposed approach and its components; section five introduces the challenges and future work; and finally, section six gives the conclusion.

II. BACKGROUND AND OVERVIEW

This section presents an overview of agile methodologies such as Scrum, Extreme programming (XP), Adaptive Software Development (ASD) and Feature-Driven Development (FDD). It also presents an overview of intelligent techniques such as fuzzy logic and multiple linear regressions.

A) Agile Methodologies

XP was proposed by Kent Beck in 2000. XP has a set of values and practices. XP includes five basic values: communication, simplicity, feedback, courage, and respect. XP supports the following practices: small iterations, code standards, spike solutions, user stories, unit testing, pair programming, and Continuous Integration [S. MERZOUK and et al,2017].

Scrum is an agile method to manage a software project. Scrum focuses on the members of staff should function to create the system flexibly in constantly changing environment [M. Al-Zawahiri and et al,2017]. It is composed of five phases as follows:

- Product backlog creation
- Sprint Planning and Sprint Backlog Creation
- Scrum Meetings
- Testing and Product Demonstration
- Retrospective and Next Sprint Planning

FDD is a reiterated software development process. Its main objective is to provide tangible, working software iteratively in a timely manner [S. MERZOUK and et al,2017]. It is composed of five activities as follows:

- Develop overall model
- Build feature list
- Plan by feature
- Design by feature
- Build by feature

ASD is an agile method for managing a large software project. It concentrates on the problems of improving complex and huge systems. This methodology highly supports incremental, iterative development, with fixed prototyping. ASD is composed of six basic characteristics: mission focused, feature based, iterative, time-boxed, risk driven and change tolerant [M. Al-Zawahiri and et al,2017].

B) Intelligent Techniques:

This section introduces overview of linear regression and fuzzy logic as follows:

1) Linear Regression Analysis

Linear regression (LR) analysis is composed of two types which are simple linear regression and multiple linear regressions. The general linear regression formula can be formulated as follows [N. R. Darwish and et al,2016]:

$$Y= \beta_0+ \beta_1x_1+ \beta_2x_2+\ldots\ldots. B_Nx_N+\epsilon \qquad (1)$$

Where:

- Y represents the dependent variable
- $x_1,x_2, ...,x_n$ are the independent variables
- β_i is the regression coefficient
- ε is the random error component.
- β_0 is the y intercept

In linear regression, there are three main outputs that include regression details, ANOVA table and coefficients table as shown in figure 1.

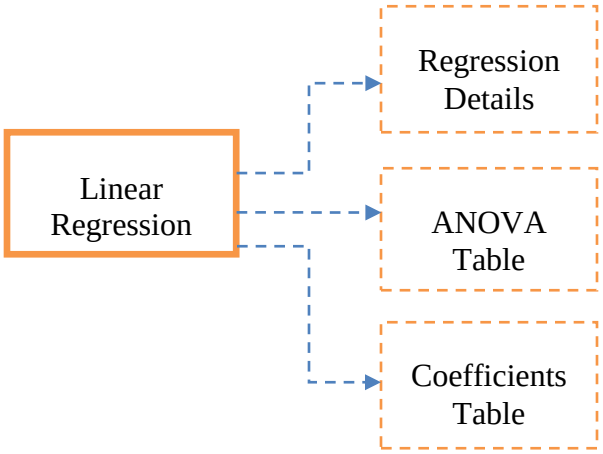


Figure.1 Three Main Results in Linear Regression

- Regression Detail shows the precise of the LR model. It is composed of three indicators which are correlation coefficient, determination coefficient, adjusted R square, and finally standard error.
- ANOVA table contains one of the most significant features which are 'significance F'. Whenever, significance F is < 0.05, this means that the proposed approach supports to define the most important factors that influence failure of agile software projects.
- Coefficients table contains one of the most significant features which are 'p-value'. If p-value in failure factors is less than 0.05, then failure factors are included in the crucial list.

2) Fuzzy Logic

Fuzzy logic is used for predicting failure of agile software projects. It is based on probabilities between 0 and 1. Fuzzy logic also contains many concepts such as linguistic variables, membership function, knowledge rules, Fuzzification and Defuzzification. These concepts will be described as follows:

- Linguistic Variables are the variables that hold items in a form of statements or sentences such as "hot", "cold" and "very cold".
- Membership functions are used to map the non-fuzzy input values to fuzzy linguistic terms and vice versa. As shown in figure 2, a membership function can has many forms such as triangular, trapezoidal, Gaussian or generalized bell.

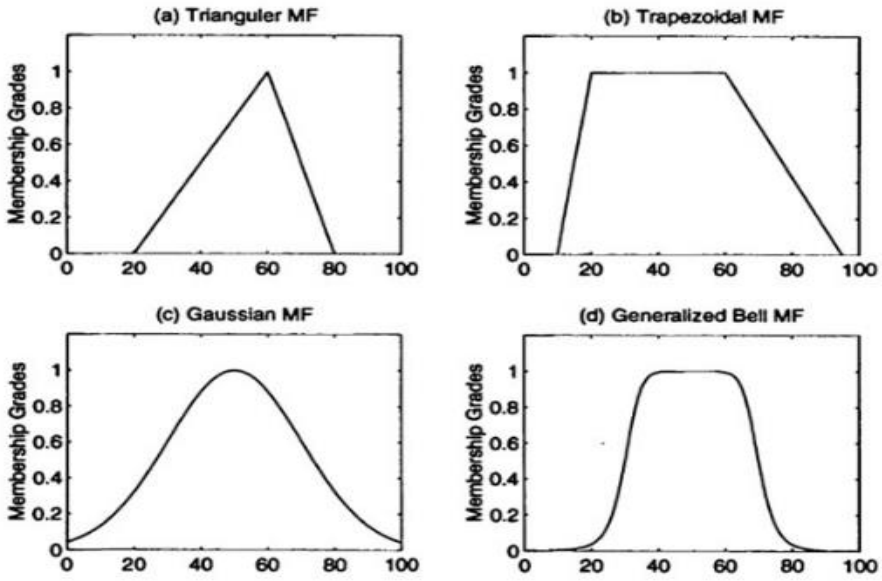


Figure.2 Sample of Different Membership Functions

- Knowledge base converts input values and output values into if-then rules for example:

If cost is low and time is high then price is Low

- Fuzzification is the process of replacing a current digit value into a fuzzy value, as shown below in figure 3.

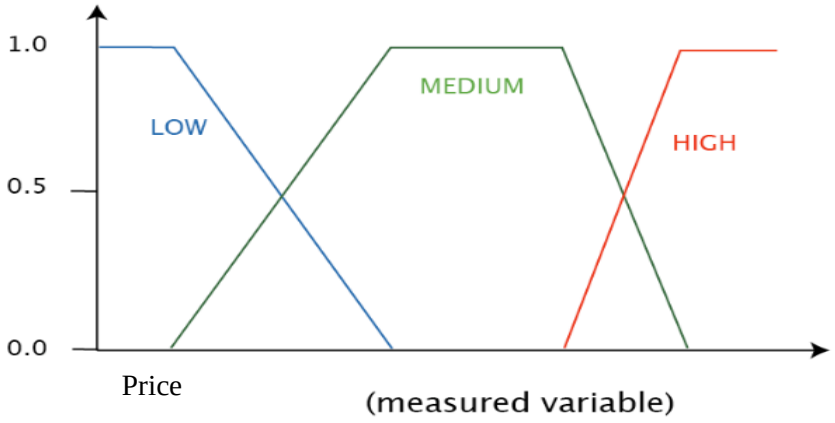


Figure.3 Example of Fuzzification process

- Defuzzification is the process in which the linguistic values are transformed into numerical values that computers can deal with it. Defuzzification (center of gravity) can be formulated as follows [A. A. Mohamed and A. S. Salama,2013, K. Jammalamadaka and V. R. Krishna,2013]:

$$Yo = (y1\ \mu Y\ (y1) + y2\ \mu Y\ (y2)) \div \mu Y\ (y1) + \mu Y\ (y2) \qquad (2)$$

Where:

- Y: The fuzzy sets to which the decision belongs.
- y1: The first decision
- y2: The second decision
- μ: Degree of membership
- Yo: The final decision

III. RELATED WORK

Through previous work, many studies were reviewed and construed, these studies utilize classical techniques such as (statistical analysis and etc.) and intelligent techniques such as (support vector machine, fuzzy logic and etc.) to detect failure and success of agile software projects, as follows:

V.Lalsing and et al, introduced a new process to assess people attributes in agile software projects based on time submission, rework standard, fault rate, and connection channels. This paper executed tests on little and moderate enterprises and using scrum method [V. Lalsing and et al,2012]. The importance of this research is to assess people attributes that influence of agile software projects by using a quantitative approach.

H. Taherdoost and et al, introduced a survey to display failure and success features of information software projects through organizational, people, process and technical features. This paper congrats on organizational features that utilize in enterprises concerned to agile methodologies [TAHERDOOST and A. KESHAVARZSALEH,2012]. The importance of this research is to use standard deviation to assess failure and success features that influence of information software projects.

M. Shepperd and et al, introduced a new approach based on unbiased statistic to detect success of agile software projects. This paper concentrates on various features such as organizational, people, process and technical features. This study seeks to detect the best method prediction of agile software projects [M. Shepperd and S. MacDonell,2012]. The importance of this research is to use unbiased statistic for predicting success of agile software projects.

S. Lee and et al, introduced a new approach to assess agile method in little projects based on agility features that are consisted of elasticity, velocity, learning and responsiveness. This paper executed tests on little projects and using scrum method [S. Lee and H. Yong,2013]. The importance of this research is to assess little projects in agile methodology by using agility features.

A. Elzamly and et al, introduced a new model based on risk features such as (determination risk, risk evaluation, risk handling, risk dominant and risk documentation) to evaluate success of agile software projects. This paper utilizes historical information from various enterprises to assess software projects [V. E. Jyothi and K. N. Rao,2011]. The importance of this research is to

use risk management equations to evaluate success of agile software projects.

M. Tanner and et al, presented a new approach to predict crucial success features to enhance successful agile adoption based on customer sharing, stakeholder sharing, team construction, project kind, and skill level of team members. This paper seeks to gather historical information through questionnaires and meeting to experts to detect the crucial success features [Tanner and U. Willingh,2014]. The importance of this research is to use mean analysis and standard deviation for determining crucial success features in agile software projects.

Feras A. Batarseh and et al, presented a modeling regression to predict failure of agile software projects. The new method seeks to define kinds of software systems failures in various companies concerned to agile methodology [Feras A. Batarseh and Avelino J. Gonzalez,2015]. The importance of this research is to use modeling regression to predict failure of agile software projects.

D. S. Nguyen introduced a new approach to define success features that influence of agile software projects based on gathering historical information from a questionnaire. This paper can define crucial success features that influence of agile software projects such as customer sharing, perfect planning and continuous delivery products [D. S. Nguyen and et al,2016]. The importance of this research is to use time series for determining crucial success features in agile software projects.

T. Chow and et al, introduced a new approach based on multiple linear regression to determine crucial success features that influence of agile software projects. This paper displays three crucial success features: (1) Delivery Strategy, (2) Agile Software Engineering Techniques and (3) Team Capability to Quality, Scope, Time, and Cost [T. Chow and D. Cao,2008]. The importance of this research is to use multiple linear regressions to determine crucial success features in agile software projects.

N. Cerpa and et al, introduced a new framework based on logistic regression to assess success features for predicting successful of agile software projects. This paper congrats on various success features such as customer sharing, perfect planning, project manager, and expansion process and development team. This paper displays that customer participation one of the most significant features that influences of agile software projects [N. Cerpa and et al,2010]. The importance of this research is to use logistic regression to predict successful of agile software projects.

R. P. Mohanty and et al, introduced a new model based on genetic algorithm (GA) to detect failure of agile software projects. This paper utilizes genetic algorithm to assess accuracy of agile software projects and detect crucial risk features [M. mayo, and S. Spacey,2013]. The importance of this research is to use genetic algorithm to predict failure of agile software projects.

D. Stankovic and et al, presented a new approach based on multiple linear regression to detect crucial success features in information technology projects. This paper displays crucial success features such as good team, Project management process and agile software engineering [D.

Stankovic and et al,2013]. The importance of this research is to use multiple linear regressions for determining crucial success features of information technology projects.

Pushpavathi T.P and et al, presented a new approach based on GA based fuzzy c-means clustering and random forest classifier to predict successful of agile software projects. This paper congrats on various success features such as Total project time and Defect counts estimation. This paper displays that the precise of the proposed approach is 93.05% [Pushpavathi T.P and et al,2014]. The importance of this research is to use genetic algorithm based fuzzy c-means clustering and random forest classifier for predicting successful of agile software projects.

H. B. Yadav and et al, presented a new method to predict successful of agile software project by using fuzzy logic. This paper executed tests on 20 projects in various sizes in small companies [H. Yadav and D. Yadav,2015]. The importance of this research is to use fuzzy logic for predicting successful of agile software projects.

T. Hovorushchenko and et al, presented a new model to predict successful of agile software project by using neural network. This paper displays the software project period and software project performance that influence of successful of agile software projects [T. Hovorushchenko and A. Krasiy,2015]. The importance of this research is to use neural network for predicting successful of agile software projects.

S.A. Rizvi and et al, presented a new method to predict early stage of successful of agile software project by using fuzzy logic. This paper displays that the precise of the proposed model is 98.4% [S. A. Rizvi and et al,2016]. The importance of this research is to use fuzzy logic for predicting early stage of successful of agile software projects.

V. Vashisht and et al, introduced a new model to predict successful of agile software project by using neuro - fuzzy. This paper displays that the precise of the proposed model is 93.4% [V. Vashisht and et al,2016]. The importance of this research is to use neuro - fuzzy model for predicting successful of agile software projects.

Through previous work, there are not official researches to define crucial failure features that effect on agile software projects. Crucial failure features of agile software projects are very significant for enterprises that combining agile method with software projects. This paper tries to detect crucial failure features of agile software projects for avoiding software projects to fail. It also aims to find the optimal intelligent techniques to predict failure of agile software projects.

IV. THE PROPOSED APPROACH

This section presents an approach to predict failure of agile software projects. The proposed approach consists of three parts: as shown below in figure 4.

1. Survey recent researches to elicit significant failure features related to agile software projects.
2. Linear regression analysis is used to define crucial failure features in agile software projects.
3. Fuzzy logic is used to predict failure of agile software projects.

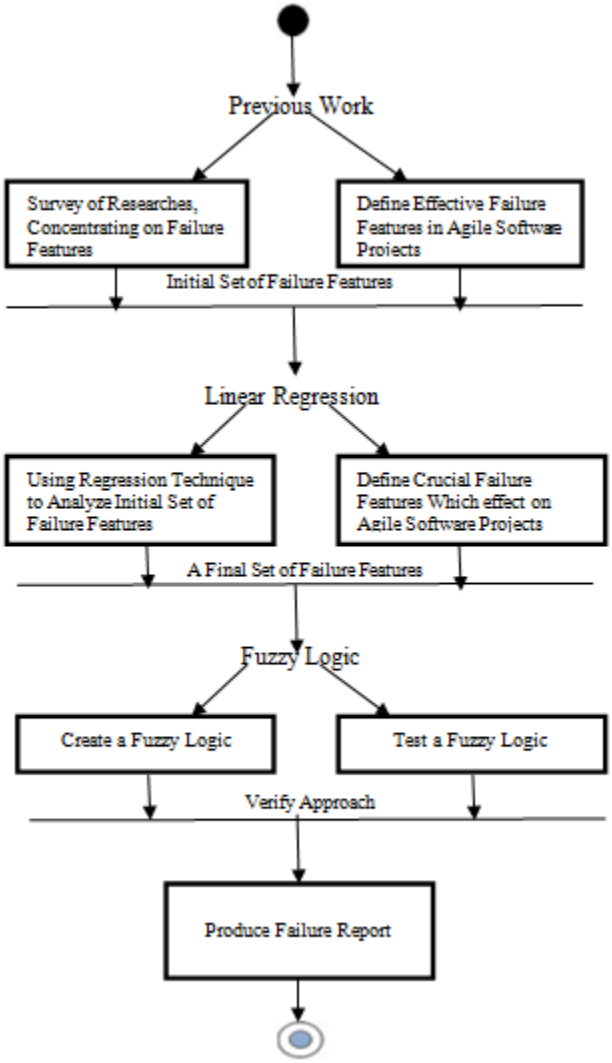


Figure.4 Proposed Approach for Predicting Failure of Agile Software Projects

Each component of the approach will be shown in the following subsections:

A. Survey of Researches

The first part is to survey recent researches for predicting failure of agile software projects. This part aims to elicit significant failure features in agile software projects. Elicit significant failure features are based on some estimation criteria (Transparency, Clarity and Easy) as shown in figure 5.

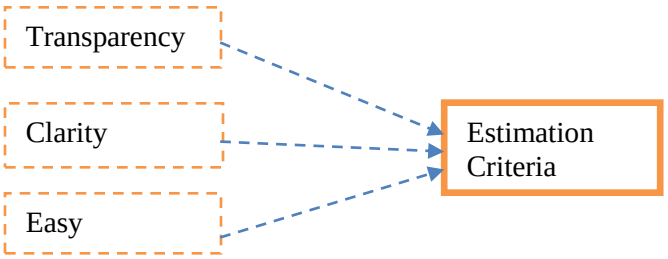


Figure.5 Estimation Criteria to Elicit Significant Failure Features

B. Linear Regression

Linear regression is used to define the crucial failure factors (CFF) in agile software projects. It consists of one dependent variable and independent variables. It is formulated as follows:

$$Y= \beta_0+ \beta_1 CFF_1+ \beta_2 CFF_2+\ldots\ldots. B_N CFF_N+\epsilon \qquad (3)$$

Where:

Y is the dependent variable (degree of influence of failure factors on agile software projects) and CFF_1, CFF_2… CFF_N are the independent variables (failure factors of agile software projects).

This section presents a general linear regression algorithm to define crucial failure factors in agile software projects.

Algorithm 1. The general Algorithm of the Linear Regression Analysis to Define Crucial Failure Features that Effect on Agile Software Projects.

1. **Input : α ⟵ (dependent variable grade of influence the failure features of agile software projects)**
2. **σ ⟵ (independent variables failure features that effect on agile software projects)**
3. **MLR = Multiple Linear Regression**
4. **PV = P-value**
5. **Build the MLR model based on the set of α and σ**
6. **Estimate the MLR model**
7. **Check the value of Standard Error.**
8. **Check the value of R- Squared.**
9. **Check the value of Adjusted R- Squared.**
10. **Check PV for each variable to define £**
11. **If (PV < 0.05)**
12. **Approve £**
13. **Else**
14. **Reject the other failure features**
15. **End If**

16. **Return £**

17. **Output: £ ⟵(a final list of crucial failure features of agile software projects)**

Linear regression is used to obtain the final list of crucial failure factors of agile software projects.

C. Fuzzy Logic

In this paper, the linguistic variables (light, medium-heavy and heavy) are used as input variables. The output variables are low, medium, high and very high. The triangular membership function is selected because it gives accurate results. As shown in figure 6, the main steps of the inference engine include: Fuzzification, knowledge base and Defuzzification.

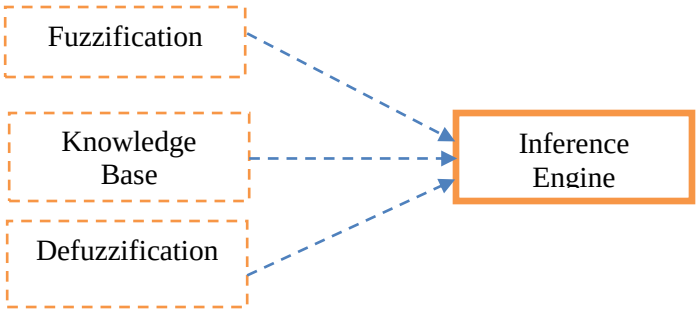


Figure.6 Three Main Steps in Inference Engine

Fuzzification is the process of replacing a current digit value (crucial failure factors in agile software projects) into a fuzzy value. Fuzzification uses linguistic values such as (light – medium-heavy and heavy) and clarifies it in membership function. Fuzzification process transforms crucial failure factors to linguistic values in membership function.

Knowledge base transforms input values (crucial failure factors in agile software projects) and output values (failure ruling in agile software projects) into if-then rules. An example of these rules is shown below:

If CFF1 is light and CFF2 is heavy then failure is Low

Defuzzification process transforms Fuzzification process (linguistic value of failure factors in agile software projects) into numerical values that the users can understand it.

This section also presents a general fuzzy logic algorithm to predict failure of agile software projects.

Algorithm 2. The general Algorithm of the Fuzzy Logic to Predict Failure of Agile Software Projects.

1. **Input : α ⟵ (a final list of crucial failure features of agile software projects)**
2. **LV = Linguistic Variables**
3. **MF = Membership Functions**
4. **F = Fuzzification**
5. **D = Defuzzification**
6. **IE = Inference Engine**
7. **Define LV in α and £**
8. **Create MF to α and £**
9. **Create the Rule Base**
10. **Transform Crisp Input Data to Fuzzy Values Using the MF (F)**
11. **Evaluate the Rules in IE**
12. **Combine the results of each rule in IE**
13. **Compute D**
14. **Check the D value for £**
15. **Return £**
16. **Output: £ ⟵ (predicting failure of agile software projects)**

V. CHALLENGES AND FUTURE WORK

Future work seeks to enhance the precise to predict failure of agile software projects by the linear regression to define the crucial features of failure and fuzzy model for predicting failure of agile software projects.

Thus, this paper will suggest a model based on hybrid intelligent techniques to predict failure of agile software projects as shown below in figure 7.

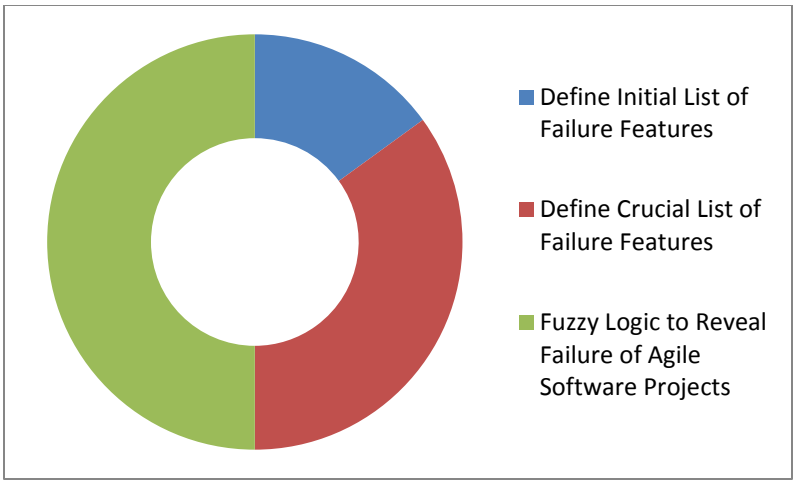


Figure.7 Future Work in Agile Software Projects

Future work seeks to realize three major targets as follows:

- Define initial list of failure features in agile software projects
- Define crucial list of failure features in agile software projects by linear regression
- Build fuzzy logic model to predict failure of agile software projects.

VI. CONCLUSION

Crucial failure factors of agile software projects are significant for software companies that are looking to adopt agile method in their software projects. In this context 17 research papers were reviewed related to predicting the failure of agile software projects and shows the advantages and disadvantage of each of them. This paper proposed a new approach for predicting the failure of agile software projects. This paper also presented intelligent techniques: linear regression and fuzzy logic. Linear regression was used to define crucial factors features in agile software projects. Fuzzy logic was used for predicting failure of agile software projects.

REFERENCES

V. Lalsing, S. Kishnah and S. Pudaruth "People Factors in Agile Software Development and Project Management", IJSEA, Vol.3, No.1, 2012, pp. 117-137.

J. Li "Agile Software Development", Technics University Berlin, Berlin, Germany, 2012.

John Hunt, "Agile software construction", Springer, 2006.

Ann L. Fruhling and Alvin E. Tarrell, "Best Practices for Implementing Agile Methods", IBM, 2008.

Laurie Williams, Robert R. Kessler, Ward Cunningham, and Ron Jeffries, "Strengthening the case for pair Programming", IEEE Software, 2000.

Mike Cohn, “Software Development using Scrum”, Addison-Wesley, 2010.

K. N. Rao, G. K. Naidu and P. Chakka “A Study of the Agile Software Development Methods, Applicability and Implications in Industry”, IJSEA, Vol.5, No.2, 2011, pp. 35-46.

N. R. Darwish “Enhancements in Scrum Framework using Extreme Programming Practices”, IJICIS, Vol.14, No.2, 2014, pp. 53-67.

Ioannis G. Stamelos and P. Sfetsos “Agile Software Development Quality Assurance”, information science references, Idea Group Inc, 2007.

V. E. Jyothi and K. N. Rao “Effective Implementation of Agile Practices”, IJACSA, Vol.2, No.3, 2011, pp. 41-48.

ELZAMLY and B. HUSSIN “An Enhancement of Framework Software Risk Management Methodology for Successful Software Development”, JATIT, Vol.62, No.2, 2014, pp. 410-423.

S. MERZOUK, S. ELHADI, H. ENNAJI, A. MARZAK and N. SAEL “A Comparative Study of Agile Methods: Towards a New Model-based Method”, IJWA, Vol.9, No.4, 2017, pp. 121-128.

M. Al-Zawahiri, M. Biltawi, W. Etaiwi, A. Shaout “Agile Software Development Methodologies: Survey of Surveys”, Journal of Computer and Communications, Vol.5, 2017, pp. 74-97.

N. R. Darwish, A. A. Mohamed and A. S. Abdelghany “A Hybrid Machine Learning Model for Selecting Suitable Requirements Elicitation Techniques”, IJCSIS, vol.14, No 6, 2016, pp. 1-12.

A. A. Mohamed and A. S. Salama “A Fuzzy Logic based Model for Predicting Commercial Banks Financial Failure”, IJCA, vol.79, No 11, 2013, pp. 16-21.

K. Jammalamadaka and V. R. Krishna “Agile Software Development and Challenges”, IJRET, Vol.2, No.8, 2013, pp. 125-129.

TAHERDOOST and A. KESHAVARZSALEH “A Theoretical Review on IT Project Success/Failure Factors and Evaluating the Associated Risks”, Mathematical and Computational Methods in Electrical Engineering, vol.1, 2012, pp. 80-88.

M. Shepperd and S. MacDonell “Evaluating prediction systems in software project estimation”, Information and Software Technology, ELSEVIER, vol.54, 2012, pp. 820-827.

S. Lee and H. Yong “Agile Software Development Framework in a Small Project Environment”, JIPS, Vol.9, No.1, 2013, pp. 69-88.

Tanner and U. Willingh “Factors Leading to the Success and Failure of Agile Projects Implemented in Traditionally Waterfall Environments”, MKL, 2014, pp. 693-701.

Feras A. Batarseh and Avelino J. Gonzalez "Predicting failures in agile software development through data analytics", Springer, 2015.

D. S. Nguyen "Success Factors That Influence Agile Software Development Project Success", ASRJETS, vol.17, No 1, 2016, pp. 172-222.

T. Chow, D. Cao "A survey study of critical success factors in agile software projects", JSS, Elsevier, vol.81, 2008, pp. 961-971.

N. Cerpa, M. Bardeen, B. Kitchenham and J. Verner "Evaluating logistic regression models to estimate software project outcomes", IST, Elsevier, vol.52, 2010, pp. 934-944.

M. mayo, and S. Spacey "Predicting Regression Test Failures Using Genetic Algorithm-Selected Dynamic Performance Analysis Metrics", Proceedings of the 5th International Symposium on Search Based Software Engineering, Springer, vol.8084, 2013, pp. 158-171.

D. Stankovic, V. Nikolicb, M. Djordjevicc, D. Caod "A survey study of critical success factors in agile software projects in former Yugoslavia IT companies", JSS, Elsevier, vol.86, 2013, pp. 1663– 1678.

Pushpavathi T.P, Suma V, and Ramaswamy V "Defect Prediction in Software Projects-Using Genetic Algorithm based Fuzzy C-Means Clustering and Random Forest Classifier", IJSER, Vol.5, 2014, pp. 888-898.

H. Yadav, D. Yadav "A fuzzy logic-based approach for phase-wise software defects Prediction using software metrics", INFSOF, 2015, vol.1, pp. 1-19.

T. Hovorushchenko and A. Krasiy "Realization of the Neural Network Model of Prediction of the Software Project Characteristics for Evaluating the Success of its Implementation", ICIDAACS, IEEE, 2015, pp. 348-353.

S. A. Rizvi, R. A. Khan and V. K. Singh "Software Reliability Prediction using Fuzzy Inference System: Early Stage Perspective", IJCA, vol.145, No 10, 2016, pp. 16-23.

V. Vashisht, M. Lal and G. S. Sureshchandar "Defect Prediction Framework Using Adaptive Neuro-Fuzzy Inference System (ANFIS) for Software Enhancement Projects", British Journal of Mathematics & Computer Science, vol.19, No 2, 2016, pp. 1-12.

A NOVEL TECHNIQUE FOR DETECTION OF SQL INJECTION

G. Yasser [2], A.A. Abd El- Aziz[1,2] ,H. Hesham [2],H. Nermin [2]

Abstract: Website applications are the important part of our daily life. Thus, the attack on those sites is rapidly increasing. One of the most famous attack techniques is the SQL injection attack (SQLIA). This attack is launched through user input developed. In addition, it runs when the user targets Web applications that use background databases. This attack changes the intentional query structure, so the retrieved data is not the required data. The best thing is that the user enters the data through the layer between the user and the database to prevent the user from entering any part of the SQL query directly to stop the attack. In this research, we suggest a new technique that acts as a layer between the user and the database. The proposed technology can be applied to new and existing websites.

Keywords: SQLI, detecting attacks, preventing attacks, static analysis and hotspot

1. Introduction

SQLIA is the most important attack where it is classified under Open Web Application Security (OWASP) Top 10 attack with high risk, impact on the server, data loss or DB corruption, irresponsibility or denial of access to DB. Can sometimes lead to a full host takeover.

Many attacks threaten the applications of the World Wide Web. The most popular and challenging attack is SQLIA. At present, most web applications are hacked using the SQLIAs method. One of the top ten security threats in web applications is SQLIA. SQLIA technologies are popular, ambitious and sophisticated. There is therefore a profound need to discover a meaningful and effective solution to this problem in the world of computer security. Discovery and prevention of SQLIAs is the subject of recent and active research in industry and academia. We suggest a new technique that acts as a layer between the user and the database. This technique will speed up SQL injection detection the following part is organized as follows: In Section 2, we introduce the SQLIA classification. The relevant work has been presented in Section 3. The proposed model is presented in Section 4. There are empirical studies in Section 5. The Comparison and results is presented in Section6, The conclusion is presented in Section 7.

2. SQLIA classification

The detailed set of each property in the SQLI attack model is the following:

2.1 Intent of attack

When the threat uses malicious SQLI input to initiate an attack, the aggressive attack is the target that the threat needs to achieve whenever the attack is successfully executed. In the following sub-sections, we explained some of the intentions of the attack.

---

1 Dept. of Information Systems, College of computer and information sciences, jouf UNI.
2 Dept. of Information Systems & Technology, Institute of Statistical Studies and Researches (ISSR), Cairo, Egypt

---

- **Determination of scalable parameters** Injectable parameters are parameters or user input fields for Web applications directly used by a server-side program to generate SQL statements that are vulnerable to SQLIA. In order to perform a successful injection, the threat agent must first detect parameters that are susceptible to SQLIA.
- **Database schema determination:** The system structure of the database is a schema. Explains relationships between fields, tables, and fields in each table. It is used by threat clients to perform a valid subsequent attack to edit data from the database.
- **Override authentication:** A mechanism used by a web application to ensure whether a user is claiming to exist. Matching the username and credentials stored in the database is the most common authentication mechanism for web applications. By passing authentication, a hacker helps impersonate another Web application user for unauthorized access.
- **Data Extraction:** Data used by Web applications are often highly sensitive to threat factors. Data acquisition attacks are the most common types of SQLIAs
- **Add or modify data:** Many features are offered to the threat by modifying the database, such that the intruder can pay much less for the online process by changing the price of the product in the database. Alternatively, thread threads can be shared in the online discussion forum by an intruder to perform subsequent cross-site scripting attacks (CSSA).
- **Implement denial of service**: These attacks aim to close the database of the application, thus refusing the service to other users. Attacks aim to lock or drop database tables also within this category.
- **Evaporation Detection**: This category refers to specific attack techniques used to avoid detection and detection by system protection mechanisms.

2.2 **Weaknesses**

Is a weakness that allows attackers to reduce system information security. Vulnerability is the intersection of three elements: system sensitivity or defect, attacker access to the defect, the attacker's ability to exploit the defect.

- **Checking Insufficient Inputs**: The input validation process is an attempt to verify or purge any specific input for malignant position. Validating insufficient input will allow the code to be implemented without proper verification of its purpose. Hackers benefit from checking for insufficient input to exploit malicious code to control attacks.
- **Premium account**: The privileged account has higher rights than any normal account. His actions can therefore be avoided from the audit and verified. Because a distinct account at the risk level, such as a supervisor account, can trigger a much more aggressive attack than the normal account, it makes it one of the most important weaknesses.

2.3 **Attack techniques**

Attack techniques are the specific means by which threat attacks are carried out using bad code. Threats may use many ways to achieve their objectives, and most of them are often mixed or used in different types.

- **Taste Science**: This technique relies on executing statements that are always correct so that queries always result in a conditional WHERE rating. Which. "Or 2 = 2" will be entered in the "Login" parameter.
- **End of line suspension:** After entering the code in a given field, the next valid symbol is skipped by using end-of-line comments. That is, "-" - is added after the input so that the rest of the queries are not treated as executable code, but comments. This is useful because threat agents may not know the syntax or fields in the server.
- **Union Inquiry:** Threats use this technique to direct servers to return data that should not be returned by developers. That is, the "UNION DROP" statement will be added, along with an additional targeted data set, so that the queries return the destination data set together with the target data set.
- **The supported query on the pig:** The threat agent may add more queries than the intended query, which means "supporting" the attack on a legitimate request. This technique relies on server configurations, which allow many different queries within one part of the code. That is, the threat factor may add a search selector such as ";", and then continue with its own order, such as "SELECT table name", which effectively determines the specified table.
- **System stored procedure**: SQLIAs of this type try to execute stored procedures in the database. Today, most database has fixed system storage procedures. Therefore, when the attacker knows which database is being used, SQLIA can perform the stored procedures that are provided by the selected database, including actions that interact with the operating system. Common misconceptions are that using stored procedures to write Web applications makes them vulnerable to SQLIAs. Developers are often surprised that their stored procedures can be vulnerable to attacks such as their regular applications.

2.4 **Countermeasures**

There are many ways in which a system administrator and developer can reject or intercept injections that have been made in their systems.

- **Query with a calculated value**: The query with the criteria considered is access to the parameterized database which is provided through the development platform such as the statement setting in SQLParameter, Java, or .NET. Better than SQL programming by string string, each parameter in a SQL query is explained by using a placeholder and the input is provided separately.
- **Less privilege**: The account used by the application to access the database must be given the minimum permissions necessary to access the objects it needs.

- **Different Accounts**: Use different accounts for different tasks you need to have a different phase of the franchise.
- **Custom error message**: Threats can lead to knowledge through extra error messages, but deleting error messages completely makes debugging a difficult task. Customizing error messages hinders the progress of the scout threat, especially in the conclusion of certain details such as SQLI injection parameters, etc.

3. Literature review

Scanning Tools Using Black Box Testing Methods These methods are two basic steps to gathering information about vulnerabilities in the Web application. The first one detects the application's workflow using a web crawler to find vulnerabilities. The second step makes the attack and monitors the application method. This technique is called the Black Box test as we do not examine the source code of the web application directly but try to generate special inputs and emulate it with this application.

Huang et al. 2003 also has proposed a black box testing technique called the WAVES scanning tool. It is also an open source crawler-based scanner supported by an analyzer engine that uses the DOM (Document Object Model) (W3C 2009) parser to provide a comprehensive description of web application components. The attacker's generator will use the crawler to inject Web application fields with a SQL injection pattern ready. The result of the attack generator depends on the Web application response and output. WAVES uses automatic learning technology to improve and improve the methodology of the attack generator.

Kals, Kirda et al. Secubat 2006 has developed an open source tool that can scan a Web application to detect vulnerabilities. Scanning tools using white frame testing tools: White box testing or static analysis methods rely on the internal code analysis of the web application and its structure to detect fragile points at the time of assembly. There have been many attempts to check the Web application to identify SQL injection vulnerabilities using the white box test method, some of which will be highlighted as follows:

Gold, Su et al. The JDBC Checker 2004, which is a statically verifiable tool to type correct queries in a SQL statement that is dynamically generated in Java, has been suggested. This method only detects security breaches for SQL-based mismatch such as incorrect logical query attacks, because it checks only the syntax of the SQL statement for errors, but SQLIA can be correctly validated and does not return database errors.

Xie, Aiken 2006 uses an analysis algorithm to analyze static open source PHP web applications for SQL injection and XSS gaps. This approach uses analysis to detect and manipulate vulnerable components of the PHP code and other scripting languages that are used to develop application pages.

The SAFELI framework is one of the white box analysis techniques proposed by **Fu (Lu et al. 2007)** for analyzing ASP.NET applications. SAFELI consists of several components; one of the main components is Microsoft Intermediate Language MSIL Instrumental, which is used to process application byte code by inserting additional functions for each access point from the application database and replacing its variants with symbolic constraints.

Randomized SQL approach: The main idea of this approach is to add numbers for each SQL keyword that is used in the application's query statement. These numbers are random numbers created randomly. Then, while executing the application, it will rewrite the SQL statements using a proxy filter and add a random number to the SQL password. Therefore, when the attacker tries to enter the application with any SQL keyword, the system will reject it because of a missing random number

(Boyd, Keromytis 2004, Kc, Keromytis et al., 2003). However, the problem with this approach is that if the attacker can determine a random number the application can be attacked.

Filter input (chain analysis) this technique relies on filtering from the input data of malicious SQL keywords that can be used to fight the database system. **Scott, et al. 2002** improved a proxy filter for the web application that could enforce the validation rule to verify user input.

Hoff Meyer, et al. 2003 developed a technique by validating input value entered by the user Compare it with data. These methods begin with a static analysis that identifies the hotspots or sensitive points in the web application and is any point that can be used by the application to access the application database. The next step is to follow the data that comes through these hot spots. Examples of these methods will be highlighted in the following:

Livshits, et al. 2005 suggested an approach to find Java Tainted Objects Instruction-Set Randomization 2004: SQLrand provides a framework that allows system developers to generate SQL queries using random keywords instead of regular keywords. The proxy between the application and the databases distorts SQL queries and defragments keywords. SQL keywords that are entered by an attacker will not be created by random keywords, and therefore the commands that are injected may result in an incorrect query. Since SQLrand uses a secret hidden key to modify the keywords, its security depends on the attackers being unable to find this key. SQLrand needs the application developer to rewrite the code.

Automated Learning Approach: **Valeur et al.2005** proposed the use of Intrusion Detection System (IDS) based on automated learning technology. IDS was well trained using a subset of application queries, building forms of queries, and then monitoring the application during operation to discover queries that did not match the form. The overall quality of IDS depends on the quality of the training package; good training does not result in a large number of negative and positive outcomes.

(Jovanovic, Kruegel et al., 2006) proposed another detection method to be implemented by Pixy

(Jovanovic, Kruegel et al. 2006), a Java-based model that can consistently analyze PHP applications. This analysis technique relies on data flow analysis to find the wrong points for a web application. However, the result of the analysis shows that there is 50% of the false positives.

Combined and Dynamic Analysis:

Huang, Yu et al. 2004 has enhanced the WebSSARI tool that uses an application-based detection algorithm to analyze the flow of application information to detect the sensitive function that can be contaminated in a PHP application.This tool has been enhanced by a run-time guard that can perform an additional examination of the inference functions in the static analysis. In addition to static analysis, a run-time guard is added which depends on the exhibits provided by the user.

Static analysis: Wassleman et al. 2004 proposed a methodology that uses a static analysis combined with automatic inference. This technique checks for the absence of any filler found in SQL queries created in the Web application. This technique is powerful only for SQLI that inserts tautology into SQL queries but cannot detect other types of SQL injection attacks.

Bohrer et al 2005. Assembling the tree SQL claim analysis to secure the weak SQL claim before and after input and allow SQL input only if there is a match in the parsing trees.They produced a study using one real Web application and applied their SQLGUARD solution to each of these applications. This has resulted in their solution preventing all SQLIA from testing them without generating false positives at all. While it stopped all SQLIAs, its solution requires the developer to rewrite all of its SQL codes to use its custom libraries, which is a roof we are trying to get rid of. However, the suggestion is not available, if the user inserts inputs that do not appear in the tree leaves.

Sue et al. 2006 Linguistic-based approach to find and stop fears of having SQLIAs by applying a tool called SQLCHECK. The author distinguishes here the parts that the user provides in queries with special characters and inflates the traditional SQL rules to the production base. The parser is created depending on the grammatical grammar. Successfully parses the generated query at run-time, if there are no SQLIAs in queries that are created after inputting user input. This method uses an encrypted key to detect user input in SQL queries. Thus, detecting the key will make the application serious. However, this method requires the application developer to rewrite the code to insert the encrypted keys into dynamically generated SQL queries.

4. PROPOSED MODEL

The recent researches minimize false positives and negatives by creating a database for storing the structures of the valid queries. In runtime validation, checking the dynamically generated query with the previously stored query structure to decide the possible SQLIA. The database of the valid query structures is made in a static analysis phase. In the static analysis phase, we are using the same technique (AMENISA) to extract the structure. Instead of using NDFA, we are storing all of the valid query structure by means of linked list illustration in which each singly link list represents a

valid query structure and to store the beginning deal with of a majority of these first link list we use a second link list referred to as main link list and each node of this list store the starting address of the singly link list. So whilst, we discovered a brand new query has arrived at the database server, we begin looking the structure of the query in our connected representation. If the search is successful seek then the query is valid otherwise, it's going to be an SQLIA.

In the proposed technique, we store the structure of the query by keeping the order of the series of token generated with the aid of the query. We are checking the sequence of token generated with the aid of the arrived query is within the same order as we saved in our legitimate query database. If the sequence of the arrived query is in equal order as the query stored in our database, the arrived query is a valid query, otherwise, if we no longer observe any ordered collection just like the arrived query in our complete database, it's a likely SQLIAs because we saved all the viable structure of the legitimate SQL query in our database in static analysis segment.

According to the proposed technique, we do not take any help from the utility software – Developed Application- so whilst a query comes from the application program for validation, it does no longer recognize that which hotspot - a hotspot is defined as a factor in the program that issues SQL queries to the underlying database,- of the application, generates this query, so we must gurantee that all of the possible structures just like the incoming query. In most cases, the incoming query is not an SQLIA, hence, due to continuous search, there is a searching trouble. To growth overall performance, we proposed a new technique to search rapidly. As there are many feasible links in the main link list which saved the beginning node of every link list storing the query structure and there is a massive request to the database server we use the looking technique in a multi-threaded manner, where in every thread is responsible for searching each individual link list stored the structure of a single query. If a thread found an appropriate course it intimates the opposite threads to terminate as is already performs a successful search. As due to the hardware steady, only few numbers of executable cores are available within the processor in order that a whole lot range of threads can't run parallel, so we should be expecting the feasible link list from among many link lists for a fast achievement. If we carefully analyse a web application maximum of the instances similar form of the query is used with different persons enter different prices and a number of the huge variety of queries few queries are frequently used. Therefore, at the same time as looking if we supply preference to paths broadly, speaking traversed then the hit ratio may be elevated. Even as begin looking, we first pick out up the often-traversed path a number of the completely feasible alternative path to discover the preferred one quickly. To do that we include successful remember at each of the nodes of the link list which factors the beginning node of some other link list which stored a legitimate query structure. While a thread performs, a hit seeks as nicely as it intimates the opposite threads to stop searching it also growth the hit count assumes that link list and ordered the complete link list that keeps the beginning node of the link lists which store the structure at the query in a descending order. to apply the threading guide, we use the JAVA Thread Pool so we are able to manage the range of threads strolling in a system. To store a valid individual query structure, we preserve the sequence of token generated by the query using a singly link list where each node store a single token of a query having an additional field to mark that the node is a position of user input or it store a token of the

static part of the query. We use a doubly linked list called as a main link list whose each node store the staring address of the a singly link list. To find an ordered sequence of token in that singly link list for an incoming query to the database we first separate the tokens from the query by a SQL parser of the specific DBMS we are trying to save from a possible SQLIA. After token separation we get the ordered sequence of the tokens of the query then we start searching the singly link list. While searching the single link list if position of the token from the incoming query matches the token of the same position in the singly link list and if it is not a consumer enter then we circulate to the following token in addition to the next node of the list until any mismatch discovered or the cease of the list. on this manner if we reach the end of the single link list and there may be no greater token left in the incoming query then it's a hit search and the incoming query is a legitimate query. Then the thread informs all other thread to forestall execution as it located a valid query structure. Then it replaces the hit count field of the node of the main link list which store the beginning address of that searched single link list and if the hit count number is greater than the hit depends on of the previous node of the main link list then it swaps the node with the preceding node. but in this matching procedure if a mismatch occurs and the node in the link list say's that it is a position of person input then we skip to the following token in the token series of the incoming query in addition to pass forward to the link list and attempt to match the token with the token in the current pointing node assuming that the previous token is a user input data value. If we found a suit at this position, we continue our matching technique or if a mismatch observed at that role then it is clear that the link list we are searching for the same query structure for the incoming query is not correct list so the thread searching this link list have to prevent its execution. The look for the similar structure for the incoming query is completed in this style in a multithreaded manner. If no comparable type of structure determined inside the dataset, then it‟s a likely SQLIA From the above description of matching method, it is clear that for a success seek, a quantity of token inside the incoming query is equal because the length of the link list saved its structure.

Figure 2 indicates a sequence of tokens extracted from an incoming query

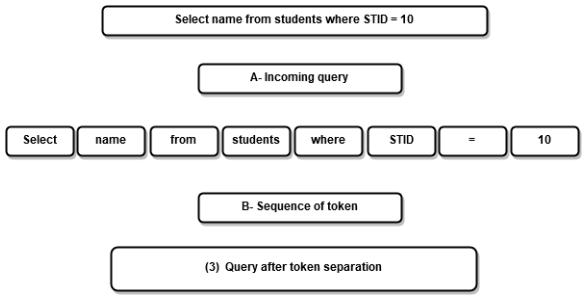


Figure 2 sequence of tokens

For runtime token matching if we used literal smart matching like others then it'll be a massive computational overhead. As an instance, if there are “n” literals are in the incoming query string and there are “q‟ query structures available within the database of the equal period of the incoming query. In worst case, if we expect the mismatch happens on the ultimate role for each and each query structure in the database and if it is an unsuccessful search then we've to check “n‟ range of literals

for every “q‟ no of the query, so the complexity might be O(n*q). So as to avoid this huge computational overhead we use a special approach in preference to the use of literal wise string matching algorithms we easy mapped each token into an integer value. We also store these integer values in our database as opposed to storing the tokens in a string layout as a query parent print. It also takes very less space for instance if there are “n‟ no of literals and “m‟ no of tokens within the query instead of string “n‟ no of values we are storing “m‟ no of values wherein m<<n. So in runtime validation the duration of a first link list storing the query having “m‟ no of tokens ism. In cases of the incoming query after token separation we remodel each token into its corresponding integer values then we begin our searching. In worst case if we assume that the mismatch occurs on the remaining function of each and each first link list and if it's far an unsuccessful seek then we've to test “m‟ no of integer values for every “q‟ no of query, so the complexity come to be O(m*q) in place of O(n*q) wherein m<<n. So that is a performance benefit. The system we are the usage of to transform a token into its corresponding integer cost is to ADD every ASCII decimal fee of a literal by its function range taking place in the token and summed it up. for example, recall the keyword “select‟ the corresponding ASCII decimal values for the literals is S=eighty-three, E=69, L=76, E=69, C=67, T=84, now the position of each literal is S=1, E=2, L=three, E=4, C=five, T=6, so after ADDING the ASCII fee of each literal with its placement we have to upload the values up. So now the fee of the token becomes 83+1+69+2+76+3+69+4+67+5+84+6=469. So the corresponding integer cost of “pick‟ is 469. Figure 3 indicates a legitimate query translation to its corresponding integer collection.

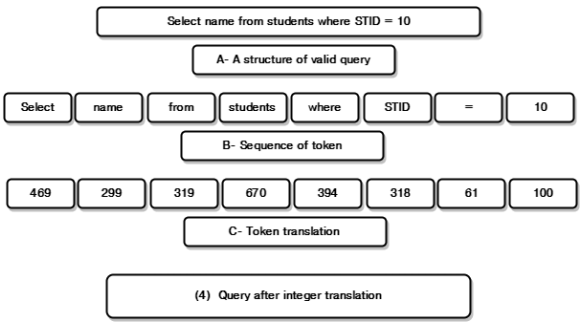


Figure 3 legitimate query translation

it's far already recognized that for a valid incoming query the range of tokens is equal as the wide variety of tokens in its corresponding query structure inside the facts base. To be able to reduce the search space we institution collectively all the query structure having same token wide variety. It manner if a query having 9 tokens belongs in a separate organization than a query having 19 tokens. So earlier than looking the same structure for an incoming query we first calculate the number of tokens it has, then we begin looking inside the group having all of the structure of legitimate query having the identical variety of tokens. To institution together all of the first link list having an equal number of tokens we use a second link list generally referred as main link list and each node holds the starting deal with of a singly link list amongst all of the singly link list having the same quantity of tokens. Meaning if we have “n‟ corporations of singly link list then we've “n‟ no of second link

list and for an incoming query, we most effective search a first second link list among the “n‟ no of the list.

Figure 4 suggests the organization representation of queries having three tokens.

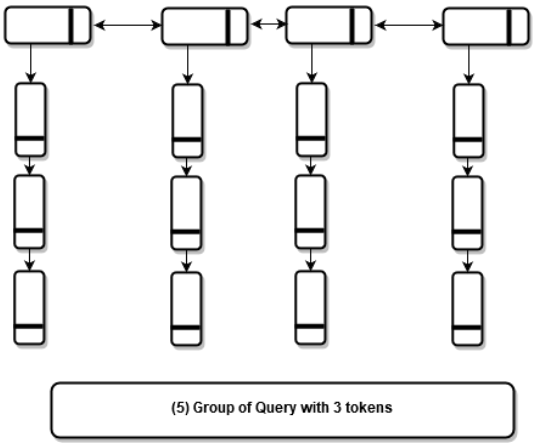


Figure 4 Group of queries

To keep the starting address of the main linked list named as the starting address of the group we use an array wherein every field of the array stores the beginning address of a collection. We store the beginning cope of a set in a way that the index of an array field need to represent the group number, the number of tokens each singly link list possesses in that institution. As an example, if a group having all the singly link list having ‘n' range of tokens then the beginning address of the group be assigned to an nth cell in the array.

although in this representation there are numerous cells within the array may additionally now not be used however through the usage of a few more storage we have a first rate gain that we don't want to search the starting cope with of a specific organization due to the fact after calculating the number of tokens in the incoming query the range itself represents the field number of the array protecting the starting address of the organization that the incoming query may additionally belong.

Figure 5 suggests the entire structure store all valid queries.

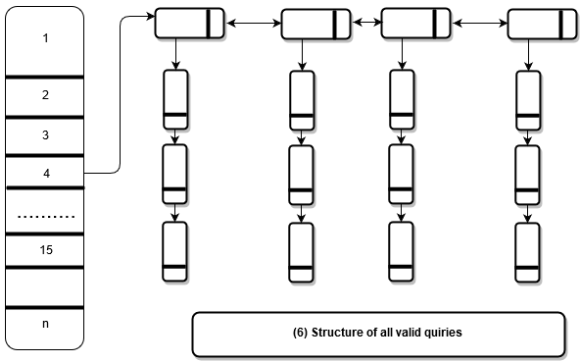


Figure 5 Structure of valid queries

5. Experimental study:

After implementing the tool to detect the timing for each of the two statements Fig 6 the result was like the following Fig 7 the used language is PHP the site used was http://roots4solutions.com/test/

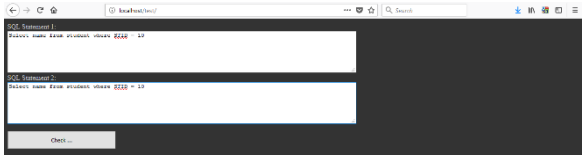


Figure 6 comparison box

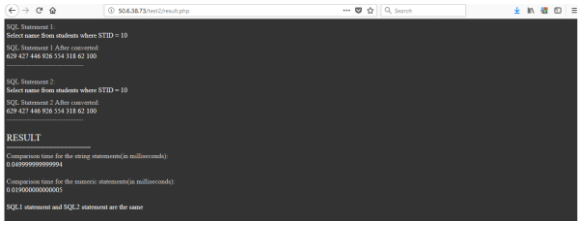


Figure 7 comparison result

Eventually after studying some pages we found out the below:

1) We need the first form of the query and the query with the injection to be in the form of tokens.

2) We have to make sure that it will take less time to be detected.

3) Sometimes Cryptography and Steganography techniques are used to improve the security.

4) Parameterized queries can be used for making dynamic queries to avoid SQL injection.

5) Every added query can be directed to the matching parser to analyze the attack and for documentation a log should be created.

6. Comparison and results:

Table I shows the comparison of our proposed model against AMNESIA according to various attack types

Detection/Prevention	Tautologies	Illegal/Incorrect	Union	Piggy Backed	Stored	Inference	Alternate	White Space
AMNESIA	Y	Y	Y	Y	N	Y	Y	ND
Proposed Model	Y	Y	Y	Y	Y	Y	Y	Y

ND in the table shows that the attack is not detected
Y in the table shows that the attack is prevented
N in the table shows that the attack is not prevented

7. Conclusion:

In this version, we analysed different SQLIAs used in the former techniques. We also discuss tactics which will be used to enhance the security and decrease the chances of getting predicted, we reduce the time to detect the injection using different techniques.

REFRENCES

BOYD, S. and KEROMYTIS, A., 2004. SQLrand: Preventing SQLIAs, Applied Cryptography and Network Security 2004, Springer, pp. 292-302.

FU, X., LU, X., PELTSVERGER, B., CHEN, S., QIAN, K. and TAO, L., 2007. A static analysis framework for detecting SQL injection vulnerabilities, Computer Software and Applications Conference, 2007. COMPSAC 2007. 31st Annual International 2007, IEEE, pp. 87-96.

F. Valeur, D. Mutz, and G. Vigna. A Learning-Based Approach to the Detection of SQL Attacks. In Proceedings of the Conference on Detection of Intrusions and Malware andVulnerability Assessment (DIMVA), pages 123–140, 2005.

G. Wassermann and Z. Su. An Analysis Framework for Security in Web Applications. In Proceedings of the FSE Workshop on Specification and Verification of Component-Based Systems (SAVCBS), pages 70–78, 2004.

G. Buehrer, Weide, and Sivilotti, "Using Parse-Tree Validation to Prevent SQLIAs," 5th International Workshop on Middleware and Software Engineering, pages 106-113, 2005

GOULD, C., SU, Z. and DEVANBU, P., 2004. JDBC checker: A static analysis tool for SQL/JDBC applications, Proceedings of the 26th International Conference on Software Engineering 2004, IEEE Computer Society, pp. 697-698

HUANG, Y.W., YU, F., HANG, C., TSAI, C.H., LEE, D.T. and KUO, S.Y., 2004. Securing web application code by static analysis and runtime protection, Proceedings of the 13th international conference on World Wide Web 2004, ACM, pp. 40-52.

HUANG, Y.W., HUANG, S.K., LIN, T.P. and TSAI, C.H., 2003. Web application security assessment by fault injection and behavior monitoring, Proceedings of the 12th international conference on World Wide Web 2003, New York, NY, USA, pp. 148-159.

HOFFMEYER, C.C. and WANG, J., 2003. Protecting Web Services from Interpretive-Language Injection Attacks.

JOVANOVIC, N., KRUEGEL, C. and KIRDA, E., 2006. Pixy: A static analysis tool for detecting web application vulnerabilities, Security and Privacy, 2006 IEEE Symposium on 2006, IEEE, pp. 6 pp.-263.

KALS, S., KIRDA, E., KRUEGEL, C. and JOVANOVIC, N., 2006. Secubat: a web vulnerability scanner, Proceedings of the 15th international conference on World Wide Web 2006, ACM, pp. 247-256.

LIVSHITS, V.B. and LAM, M.S., 2005. Finding security vulnerabilities in Java applications with static analysis, Proceedings of the 14th conference on USENIX Security Symposium 2005, pp. 18-18.

M. Muthuprasanna, Ke Wei, Suraj Kothari. Eliminating SQLIAs. Eighth IEEE International Symposium on Web Site Evolution,pages 22-32,Sept,2006.

OWASP Top 10 Web Application Vulnerabilities, https://www.owasp.org/index.php/Category:OWASP_Top_Ten_Project

S. Boyd and A. Keromytis. SQLrand: Preventing SQLIAs. In Proceedings of the Applied Cryptography and Network Security (ACNS), pages 292–304, 2004.

SCOTT, D. and SHARP, R., 2002. Abstracting application-level web security, Proceedings of the 11th international conference on World Wide Web 2002, Citeseer, pp. 396-407.

The Three Tenets of Cyber Security". U.S. Air Force Software Protection Initiative. Retrieved 2009-12-15.

W. G. J. Halfond, et al., "A Classification of SQLIA and Countermeasures," in the IEEE International Symposium on (SSE) Secure Software Engineering, Arlington, USA,2006.

W. Halfond and A. Orso. AMNESIA. In Proceedings of the 20th IEEE International Conference, pages 174–183, 2005.

XIE, Y. and AIKEN, A., 2006. Static detection of security vulnerabilities in scripting languages, Proceedings of the 15th conference on USENIX Security Symposium 2006, pp. 179-192.

Z. Su and G. Wassermann. Annual Symposium on (POPL) Principles of Programming Languages, pages 372–382, 2006.

جامعة القاهرة
معهد الدراسات والبحوث الإحصائية

المؤتمر السنوى الثالث والخمسين
للإحصاء وعلوم الحاسب
وبحوث العمليات

نظم المعلومات

3-5 ديسمبر 2018

Cairo University
Institute of Statistical Studies and Research

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

Information Systems

3-5, Dec, 2018