Cairo University

Institute of Statistical Studies and Research

The 52nd Annual Conference on Statistics, Computer Sciences and Operations Research

Computer Science

25-27 Dec. 2017

Index

Computer Sciences

[1] Dept. of computer sciences Institute of statistical studies and researches ,Cairo University.

[2] Dept. of computer sciences Institute of statistical studies and researches ,Cairo University.

The 52^{nd} Annual Conference on Statistics, Computer Sciences and Operation Research 25–27 Dec,2017

An Improvement and Implementation of the DPLL Satisfiability Algorithm

Munira A. El-Maksoud , Areeg Abdalla *

Abstract

This paper introduces a modification of the well known Satisfiability solver, MINISAT. A new measure of the activity of the variable, determined by its occurrences in the not yet satisfied clauses, is introduced. Hence, variable activities are changed dynamically (increased and decreased) by adding and removing clauses. It is applied on an interesting mathematical problem, in finding the van der Waerden numbers, which are known to be very difficult to compute.

keywords: Satisfiability, DPLL, MINISAT, Branching Rules, van der Waerden numbers.

1 Introduction

Boolean Satisfiability Problem (SAT) is the problem of deciding whether a given boolean formula is satisfiable or not. Since SAT is very important in many applications, there has been many algorithms for testing the satisfiability. The most well-known one is introduced in 1962 by " M. Davis, H. Putnam, G. Logemann and D. Loveland "(DPLL [6]). It is considered the basis for almost all modern SAT solvers. It is known that SAT can be used (utilized) for solving problems of various applications after encoding them into SAT. For mention one example, some of mathematical problems can be encoded as SAT then solved using SAT solvers. A SAT solver(based on DPLL [6]) is a software, and many of the SAT solvers are written in C or C^{++} [25]. One may design his (or her) own SAT solver for a specific application, and in this case he (or she) has to choose between two ways, either to build their solver from scratch which may be hard work and time consuming

*Department of Mathematics, Faculty of Science, Cairo University.

or to modify an existing solver to meet the requirements (goals). For example, one can create new ideas in the branching rules resulting in various versions of already existing SAT solvers.
In this research we choose to modify the MINISAT solver in terms of calculation method of the activity of a variable and apply this modification in finding the van der Waerden numbers as a mathematical application. The rest of this paper is organized as follows: section2 concerns the Satisfiability Problem (SAT), the relation between SAT and combinatorics and introduces the problem of van der Waerden numbers as an example of this relation. Section3 presents a comparison between the results of both of the solvers VANSAT, MINISAT on a number of examples. Section4 summarizes the work we have done and what we intend to in the future.

2 SAT and Combinatorics

Areas of satisfiability and combinatorics may help advance each other. On one hand, thanks to the significant efficiencies of modern SAT solvers it became possible to encode many of combinatorics problems as boolean formulas and then solve their corresponding satisfiability problems. In this scenario, novel results in combinatorics are obtained. On the other hand, combinatorics problems can be utilized as a rich source of structured formulas for developing new generations of SAT solvers [7].

2.1 The Satisfiability Problem (SAT)

SAT is a central problem in theoretical computer science studied by computer scientists and mathematicians, which can be formulated as [1]:
A *truth assignment* is a mapping f that assigns 0 (interpreted as False) or 1 (interpreted as True) to each variable in its domain, let us enumerate all the variables in the domain as $x_1, x_2, \ldots, x_n$. The complement $\overline{x}_i$ of each variable x_i is defined by: $f(\overline{x}_i) = 1 - f(x_i)$ for all truth assignments *f*.
Both x_i and $\overline{x}_i$ are called *literals*, if $u = \overline{x}_i$ then $\overline{u} = x_i$. A *clause* is a set of (distinct) literals, and a *formula* is a family of (not necessarily distinct) clauses. For example, $\{x_1, \overline{x}_2, x_3\}$ is a clause with three distinct literals and $\{ \{x_1, x_2\}, \{x_1, \overline{x}_2\}, \{\overline{x}_1, x_2\}, \{\overline{x}_1, \overline{x}_2\} \}$ is a formula with four clauses over two variables.
A truth assignment *satisfies* a clause if it maps at least one of its literals to 1, the assignment *satisfies* a formula if and only if it satisfies each of its clauses. A formula is called *satisfiable* if it is satisfied by at least one truth assignment, otherwise it is called *unsatisfiable*. The problem of recognizing

satisfiable formulas is known as the *satisfiability problem*, or SAT for short. A SAT *instance* is any formula for which we want to solve the SAT problem.

2.2 The DPLL Algorithm

DPLL [6] was introduced in 1962 as a refinement of its earlier M. Davis, H. Putnam (DP) algorithm. Essentially, it is a (complete - "depth-first" - backtracking) search algorithm. More recently, João P. Marques-Silva and Karem A. Sakallah introduced Generic seaRch Algorithm for the Satisfiability Problem (GRASP) [19] as an extension of DPLL [6] with learning and non-chronological backtracking. In recent decades, GRASP prompts research on conflict-driven clause learning (CDCL) solvers. Most state-of-the-art SAT solvers as Glucose, CryptoMiniSat and MINISAT follow CDCL.

2.3 Branching Rules

The rule by which a solver selects an unassigned (free) variable and assigns it a value (branching on) is called a branching rule. It is known that choosing a suitable branching rule is an important since it affects the performance. Selecting different branching rules for the same algorithm may produce search trees with quite different sizes. Over the years, various rules have been proposed by researchers who try to find new ideas for improving the performance of SAT solvers. Some examples are Jeroslow-Wang [13], BOHM's [5] and MOMS [10].

Here, we can talk about two kinds of branching rules:
State-dependent and Variable State Independent Decaying Sum (VSIDS) rules.

State-dependent rules Where the occurrences of literals differ by different assignments of the variables. Hence to maintain the occurrences, the solver should pay a high price where occurrences should be recalculated each time a solver assigns (or un-assigns) a variable. Some examples are Maximum Occurrence in clauses of Minimum Size (MOMS) [10] and Dynamic Largest Individual Sum (DLIS).

Variable State Independent Decaying Sum (VSIDS) Because of the overhead in the solver for maintaining the literals occurrences in State-dependent strategy, new more efficient branching rules are needed. Chaff [20] proposed VSIDS which imposes an order on literals, where each literal has a score related to its occurrences. The activity of the

variables is captured by the literals scores which do not depend on assignments. For branching, VSIDS will select the (unassigned) variable corresponding to the literal possessing the highest score. Periodically the scores of all literals are multiplied by a factor less than one, hence, decaying .

" More recently, BerkMin [11] proposed a new strategy " pushes the idea of VSIDS further ", where the activity is captured by conflicts. More clearly, after each conflict, a learned clause is generated through a process called " resolution process " and as a result, the scores of all literals in the clauses involved in this resolution have to be increased. BerkMin also decayed the scores periodically.

2.4 Sat Solvers

The possibility of encoding many of practical problems in diverse fields (like software verification [12], circuit testing [23], AI planning [17]) as SAT instances was the motivation for the research in developing new SAT solvers with significant efficiencies. Some examples are multi-SAT [21], Glucose and Syrup in the SAT'17 [4], Nigma [14](and its improved versions Nigma 1.1 [15], Nigma 1.2 [16]) and Glulu [24]. The following is a comparison between the two solvers MINISAT, VANSAT(VAN der Waerden numbers SAT).
MINISAT, MINImal SAT-solver, is a CDCL (CHAFF [20] - based) SAT solver written by Eén and Sörensson. MINISAT makes no distinction between the two phases of a variable. It attaches each variable with an activity which is increased whenever the variable appears in a conflict clause(bumping). MINISAT bumps variables with larger and larger numbers until reaching a limit (predefined number), at that point all variables activities are scaled down. MINISAT uses a heap to sort the variables by the activity at all times and for branching, it selects the variable with the highest activity [9].
On another side, VANSAT is a modification of MINISAT1.14 where the activity of a variable is captured by its occurrences in the not satisfied clauses. Hence, activities are changed dynamically(increased and decreased) by adding and removing clauses.
More accurately, the strategy of VANSAT is:
I- increasing the activity of all variables that appear in each new added clause (learnt or problem clause).
II- decreasing the activity of all variables that were appearing in each deleted clause (where deletion of clauses occur in many situations).

2.5 SAT and van der Waerden numbers

The problem of computing van der Waerden numbers is one of the most interesting examples in combinatorics which can be encoded as formulas. These problems can be represented by parameterized formulas in such a way that decisions concerning the satisfiability of these formulas determine the van der Waerden numbers in question. The next two sub-subsections define the van der Waerden numbers and discuss the SAT Encoding of them.

2.5.1 van der Waerden numbers

The van der Waerden number $\mathrm{w}(r;\mathrm{t}_1,\mathrm{t}_2,\ldots,\mathrm{t}_r)$ is the least integer m such that for every partition $C_1 \bigcup C_2 \bigcup \ldots \bigcup C_r$ of the set $\{1, 2, \ldots, m\}$, there is an index j in $\{1, 2, \ldots, r\}$ such that C_j contains an arithmetic progression(AP) of t_j terms [2].
where, $r > 0$ is the number of the blocks of the partition, t_j's are the lengths of the AP's, C_j's are the blocks of the partition and arithmetic progression(AP) is a sequence of numbers such that the difference between the consecutive terms, d, is constant. For example, the sequence 5, 7, 9, 11, 13, 15,... is an arithmetic progression with common difference(d) of 2.
Finding the value of van der Waerden numbers presents a challenging problem, since the underlying principle behind their computation is still unknown. The interest of many researchers in van der Waerden numbers was the reason behind the computing of many of them, Table 1 lists some of the known van der Waerden numbers:

2.5.2 SAT Encoding of van der Waerden numbers

Given positive integers r,t_1,t_2,...,t_r and n, we construct a SAT formula,F, which is satisfiable if and only if $n <\mathrm{w}(r;\mathrm{t}_1,\mathrm{t}_2,\ldots,\mathrm{t}_r)$.

To compute the van der Waerden numbers we use the following algorithm: For consecutive integers $m = r+1,\ r+2,\ \ldots$ we test whether the formula F is satisfiable. If so, we continue. If not, we return m and terminate the algorithm. (note: $\mathrm{t}_1,\ \mathrm{t}_2,\ \ldots,\ \mathrm{t}_r \geq 2$ and then $m > r$)

Consider the following two cases :
(I) $r = 2$, we have n variables, x_1,x_2,...,x_n and two types of clauses:
(1) $\{\overline{x}_a,\ \overline{x}_{a+d}\ ,\ \ldots,\ \overline{x}_{a+d(t_1-1)}\ \}$ with $a \geq 1$, $d \geq 1$, $a + d(t_1 - 1) \leq n$
(2) $\{x_a,\ x_{a+d}\ ,\ \ldots,\ x_{a+d(t_2-1)}\ \}$ with $a \geq 1$, $d \geq 1$, $a + d(t_2 - 1) \leq n$

Table 1: SOME OF THE KNOWN VAN DER WAERDEN NUMBERS

w(r;t_1,t_2,...,t_r)		REFERENCE
w(2; 3, 11)	114	LANDMAN, ROBERTSON AND CULVER [18]
w(2; 3, 13)	160	LANDMAN, ROBERTSON AND CULVER [18]
w(3; 2, 3, 7)	55	LANDMAN, ROBERTSON AND CULVER [18]
w(3; 3, 4, 4)	89	LANDMAN, ROBERTSON AND CULVER [18]
w(4; 2, 2, 3, 6)	48	LANDMAN, ROBERTSON AND CULVER [18]
w(4; 2, 2, 3, 7)	65	LANDMAN, ROBERTSON AND CULVER [18]
w(5; 2, 2, 2, 3, 3)	20	LANDMAN, ROBERTSON AND CULVER [18]
w(5; 2, 2, 3, 3, 3)	41	LANDMAN, ROBERTSON AND CULVER [18]
w(6; 2, 2, 2, 2, 4, 4)	56	Tanbir Ahmed [1]
w(6; 2, 2, 2, 3, 3, 3)	42	Tanbir Ahmed [1]
w(7; 2, 2, 2, 2, 2, 3, 3)	24	Tanbir Ahmed [1]
w(7; 2, 2, 2, 2, 2, 3, 4)	36	Tanbir Ahmed [1]
w(8; 2, 2, 2, 2, 2, 2, 3, 3)	25	Tanbir Ahmed [1]
w(9; 2, 2, 2, 2, 2, 2, 2, 3, 3)	28	Tanbir Ahmed [1]

Where x_i = TRUE encodes $i \in C_1$ and x_i = FALSE encodes $i \in C_2$.
Clauses (1) prevent the existence of an arithmetic progression of length t_1 in C_1 and Clauses (2) prevent the existence of an arithmetic progression of length t_2 in C_2 .

(II) $r > 2$, we have nr variables, $x_{i,j}$'s , with i=1,2,...,n and j=1,2,...,r Where the variable $x_{i,j}$ takes the value TRUE if and only if the integer i belongs to a block C_j of a partition. Here , we have three types of clauses:

(1) $\{x_{i,1}, x_{i,2}, \ldots, x_{i,r}\}$, for each integer i , to ensure that integer i belongs to at least one block of the partition.

(2) $\{\overline{x}_{a,j}, \overline{x}_{a+d,j}, \ldots, \overline{x}_{a+d(t_j-1),j}\}$, for $1 \leq j \leq r$, $1 \leq a \leq n - t_j + 1$ and $1 \leq d \leq \lfloor (n-a)/(t_j - 1) \rfloor$, to ensure that no arithmetic progression of length t_j in block C_j .

(3)$\{\overline{x}_{i,s}, \overline{x}_{i,t}\}$, for $1 \leq i \leq n$, $1 \leq s < t \leq r$ to ensure that integer i belongs to at most one block of the partition.

3 Experimental work and Results

The experimental work was as follows, first, since most of the known branching rules take into account the occurrences of the variables in the clauses that have not been satisfied we computed the occurrences of the variables in the formulas . Codes are written to find the variables with the maximum number of occurrences based on the above SAT encoding for the van der Waerden numbers. The output of the programs indicated a *symmetry* for the number of occurrences of the variables in the clauses. Second, we wrote a program to generate the instances in DIMACS cnf format since the majority of SAT solvers, including MINISAT, accept the input in this format. Finally, to implement a new sat solver based on MINISAT, it is required first to install its source-code from the MINISAT web-page and then editing it.

The following tables(except the last) show a comparison between results of both of VANSAT and MINISAT Solvers for a number of examples of van der Waerden numbers in terms of the number of Restarts, Conflicts, Decisions, Propagations, Conflict literals, Memory used and Cpu Time where VANSAT works better. The last table shows a comparison between the two solvers for a Random-3-SAT instance.

Table 2 shows a comparison between VANSAT and MINISAT in finding the van der Waerden number w(3;2,3,3) for the integer n=13. It is clear that the VANSAT gives much better results in the number of Conflicts, Decisions, Propagations and Conflict literals. It also used less Memory.

Table 2: van der Waerden number : w(3;2,3,3) for n=13

EXAMPLE1:w(3;2,3,3), n=13	MINISAT	VANSAT
Restarts	1	1
Conflicts	18	1
Decisions	25	6
Propagations	274	51
Conflict literals	132	7
Memory used	5.81 MB	4.81 MB
Cpu Time	0 s	0 s

Table 3 shows a comparison between VANSAT and MINISAT in finding the van der Waerden number w(3;2,3,3) for the integer n=14 , where VANSAT works better than MINISAT and gives less number of Conflicts, Decisions,and much better in Propagations, and Conflict literals. It also used less Memory.

Table 3: van der Waerden number : w(3;2,3,3) for n=14

EXAMPLE2:w(3;2,3,3), n=14	MINISAT	VANSAT
Restarts	1	1
Conflicts	67	61
Decisions	76	70
Propagations	1101	985
Conflict literals	387	229
Memory used	5.81 MB	4.88 MB
Cpu Time	0 s	0 s

Table 4 shows a comparison between VANSAT and MINISAT in finding the van der Waerden number w(3;2,3,3) for the integer n=8 , though the VANSAT works as good as the MINISAT, it used less Memory.

Table 4: van der Waerden number : w(3;2,3,3) for n=8

EXAMPLE3:w(3;2,3,3), n=8	MINISAT	VANSAT
Restarts	1	1
Conflicts	0	0
Decisions	7	7
Propagations	24	24
Conflict literals	0	0
Memory used	5.81 MB	4.81 MB
Cpu Time	0 s	0 s

Table 5 shows a comparison between VANSAT and MINISAT in finding the van der Waerden number w(3;2,3,5) for the integer n=20 , where VANSAT works better than MINISAT and gives less number of Conflicts, Propagations and Conflict literals . It also used less Memory.

Table 5: van der Waerden number : w(3;2,3,5) for n=20

EXAMPLE4:w(3;2,3,5), n=20	MINISAT	VANSAT
Restarts	1	1
Conflicts	4	0
Decisions	14	19
Propagations	141	60
Conflict literals	64	0
Memory used	5.81 MB	4.81 MB
Cpu Time	0 s	0 s

Table 6 shows a comparison between VANSAT and MINISAT in finding the van der Waerden number w(4;2,2,3,3) for the integer n=16 , where VANSAT works much better than MINISAT and gives less number of Restarts, Conflicts, Decisions, Propagations and Conflict literals . It also used less Memory and was better in CPU usage.

Table 6: van der Waerden number : w(4;2,2,3,3) for n=16

EXAMPLE5:w(4;2,2,3,3), n=16	MINISAT	VANSAT
Restarts	3	1
Conflicts	255	58
Decisions	367	263
Propagations	5149	1369
Conflict literals	2810	319
Memory used	5.81 MB	4.81 MB
Cpu Time	0.015 s	0 s

Table 7 shows a comparison between VANSAT and MINISAT on Random-3-SAT instance for clause length=3, 50vars and 218clauses, where VANSAT works better than MINISAT and gives less number of Conflicts, Decisions, Propagations and Conflict literals . It also used less Memory and was better in CPU usage.

Table 7: Random-3-SAT : clause length=3, 50vars, 218clauses

EXAMPLE6:Random-3-SAT	MINISAT	VANSAT
Restarts	1	1
Conflicts	44	35
Decisions	58	44
Propagations	594	501
Conflict literals	130	107
Memory used	5.81 MB	4.81 MB
Cpu Time	0.015 s	0 s

4 Conclusion and Future Work

This paper concerned the Satisfiability Problem (SAT) as a central problem in theoretical computer science and its well known DPLL algorithm. It introduced the solver VANSAT as a modification of the MINISAT solver(which based on DPLL). In other words, the paper introduced an implementation of DPLL different from that of the MINISAT where a new measurement of the activity of a variable is considered. Experimental results proved that VANSAT worked better in finding some of van der Waerden numbers in terms of the number of Restarts, the number of Conflicts, the number of Decisions, the number of Propagations, the number of Conflict literals, Memory usage and CPU Time.
We intend to study the symmetry that has been found in the occurrences of the variables during implementation. We also plan to find new van der Waerden numbers.

References

[1] Tanbir Ahmed, *Some new van der Waerden numbers and some van der Waerden-type numbers*, Integers 9, A06, 65–76. MR2506138, 2009.

[2] Tanbir Ahmed, *Two new van der Waerden numbers: w(2; 3, 17) and w(2; 3, 18)*, Integers, 10, A32, 369–377, 2010.

[3] Gilles Audemard and Laurent Simon. Glucose http://www.labri.fr/perso/lsimon/glucose

[4] Gilles Audemard and Laurent Simon. *Glucose and Syrup in the SAT'17*. In Proceedings of SAT Competition 2017 , Solver and Benchmark Descriptions, pp.16–17, 2017.

[5] M. Buro, H. Kleine Büning. *Report on a SAT competition*. Technical report, University of Paderborn, November 1992.

[6] Martin Davis, George Logemann and Donald Loveland. *A machine program for theorem-proving*, Communications of the ACM,5(7):394–397, MR0149690, 1962.

[7] Michael R. Dransfield, Lengning Liu. *Satisfiability and computing van der Waerden numbers*. The ELECTRONIC JOURNAL OF COMBINATORICS 11, 2004.

[8] Niklas Eén, Niklas Sörensson. MINISAT http://minisat.se

[9] Niklas Eén, Niklas Sörensson. *An Extensible SAT-solver*. In Theory and Applications of Satisfiability Testing, pp. 333–336. Springer, 2004.

[10] Jon William Freeman. *Improvements to Propositional Satisfiability Search Algorithms*,PhD thesis, Departement of computer and Information science, University of Pennsylvania, Philadelphia, 1995.

[11] Eugene Goldberg, Yakov Novikov. *BerkMin: A fast and robust SAT-solver*, In Proceedings of the Conference on Design Automation and Test in Europe pp. 142-149 Paris France Mar. 2002.

[12] Daniel Jackson and Mandana Vaziri. *Finding Bugs with a Constraint Solver*, Proc. International Symposium on Software Testing and Analysis(ISSTA?00), Portland, Oregon, August 2000.

[13] Robert G. Jeroslow and Jinchang Wang *Solving propositional satisfiability problems*. Annals of Mathematics and Artificial Intelligence,Volume 1 pp.167–187, 1990.

[14] Chuan Jiang and Ting Zhang. *Nigma: A Partial Backtracking SAT Solver*. In Proceedings of SAT Competition 2013 , Solver and Benchmark Descriptions, pp. 62–63 , 2013.

[15] Chuan Jiang and Gianfranco Ciardo *Nigma 1.1*. In Proceedings of SAT Competition 2014 , Solver and Benchmark Descriptions, p. 53 ,2014.

[16] Chuan Jiang and Gianfranco Ciardo *Nigma 1.2*. at SAT Race 2015, 2015.

[17] Henry Kautz and Bart Selman.*Planning as Satisfiability*. In European Conference on Artificial Intelligence, volume 92, pp. 359–363, 1992.

[18] Bruce Landman, Aaron Robertson, Clay Culver. Some new exact van der Waerden numbers, Integers: ELECTRONIC JOURNAL OF COMBINATORIAL NUMBER THEORY, 5(2) (2005), A10, MR2192088.

[19] João P. Marques-Silva, Karem A. Sakallah. *GRASP: A New Search Algorithm for Satisfiability*, IEEE Transactions on Computers, 48: 506–521, 1999.

[20] Matthew W. Moskewicz, Conor F. Madigan, Ying Zhao, Lintao Zhang, and Sharad Malik. *Chaff: Engineering an Efficient SAT Solver*. In Proceedings of the 38th conference on Design Automation, pp. 530–535, New York,USA, 2001.

[21] Sajjad Siddiqi and Jinbo Huang. *multi-SAT: An Adaptive SAT Solver* . In Proceedings of SAT Competition 2016 , Solver and Benchmark Descriptions, p.54 , 2016.

[22] Mate Soos. Cryptominisat http://www.msoos.org/cryptominisat2

[23] Paul R. Stephan, Robert K. Brayton, Alberto L.Sangiovanni-Vincentelli. *Combinational Test Generation Using Satisfiability*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 15, pp. 1167-1176, 1996.

[24] Aolong Zha. *Glulu*. In Proceedings of SAT Competition 2017 , Solver and Benchmark Descriptions, p. 18 , 2017.

[25] Lintao Zhang, Sharad Malik. *The Quest for Efficient Boolean Satisfiability Solvers*, In Proceedings of the 14th International Conference on Computer Aided Verification . Springer-Verlag, July 2002.

A proposed Architecture for Shopping Association Mining

Essam Eldin Mosallam[1], Reda Abd Elwahab,Alkhoribi[2,] Mohamed Ragaie Sayed Osman[3]

Abstract

This paper presents Mobil application for association mining of shopping based on Apriori algorithm. The proposed system shows and displays some offers and deals from various branches. The system provides the analytics for the seller; as the demand for some varieties and their weakness in other varieties and the whole application is organized on the Cloud. The architecture includes three levels; the front-end, middle and back-end. The front-end level of the site-based Mobile shopping application is made up of Android Mobile devices, to buy miscellaneous products from various nearby branches. The front-end level also displays the link between items purchased. The middle repository level provides a Web service to generate returned from relational database. The Exchanged information and data between application and servers is stored in the Cloud. The background level offers a Web server and a MySQL database. In this paper, we propose a architecture that reduce the communication overhead in existing Mobile Agent-based Distributed Association Rule Mining (MAD-ARM).

Keywords: Cloud, MCC, SaaS, Market Basket Analysis (MBA) and (MAD-ARM).

1. INTRODUCTION

We know that mobile devices are constrained by their processing power, battery life and storage. However, cloud computing provides an illusion of infinite computing resources. Mobile cloud computing is a new platform combining the mobile devices and cloud computing to create a new infrastructure, whereas cloud performs the heavy lifting of computing-intensive tasks and storing massive amounts of data. In this new architecture, data processing and data storage happen outside of mobile devices. Cloud computing combined with data mining can provide powerful

[1] Master student in Arabic Academy for Science, Technology and Maritime Transport.
[2] Vice Dean for Educational and Student Affairs, Faculty of Computers and Information, Cairo University.
[3] Assistant Professor, Arabic Academy for Science, Technology and Maritime Transport, College of Computing and Information Technology

capacities of storage and computing and an excellent resource management. Data mining in the cloud computing environment can be considered as the future of data mining because of the advantages of cloud computing paradigm. Mobile cloud computing (MCC) is a new emerging research field. Today's mobile devices have many advanced features such as mobility, communication, and sensing capabilities, and can serve as the personal information gateway for mobile users. However, when running complex data mining and storing operations, the computation, energy, and storage limitations of mobile devices demand an integrated solution relying on cloud-based computation and storage support. The Mobile shopping application consists of Mobile devices which limited memory and processing speed. The Cloud, on the other hand, provides a large storage and speed for data stores. The Mobile Cloud Computing (MCC) is an area where three things are involved: Mobile devices, networking, and the Cloud. Data is stored and processed on the Mobile devices on the cloud using a giant computer, and the results then are displayed through output devices such as the monitor. Google Drive, Gmail, Google Drive and Google Maps are already used as examples of Mobile cloud computing. Thus, Mobile cloud computing covers the disadvantages and weaknesses of Mobile devices such as short storage area and processing power. [1] The research is implemented in the Cloud for supermarket shopping products. Graphical user interface (GUI) was designed by using location-based services and association rule mining. This research covers these drawbacks and implements association rule mining on the data gathered from the Mobile application. [2]

- **Objective**

This paper is specially designed for product purchasing in the city for malls and a local market nearby the customer location. It uses Wi-Fi Global Positioning System (GPS) and Mobile network to get the current position of the customer and displays any registered supermarket's branch on the server from customer's location. Association rule mining as a technique of data mining is used to find the offers associated with products. Technically, data mining is the process of extraction of interesting information or patterns from data in the large database. Association rule mining is widely used in market basket analysis. This method benefits retailers in numerous ways for marketing or planning shelf space.

The paper organization

The paper is organized as follows:

2. Background: review the Mobile Cloud Computing (MCC), Market Basket Analysis (MBA), Association Rule Mining and Location-based services
3. Proposed Architecture: review the proposed, system architecture, software and data Set.
4. Related Work: reviews the work previously done by researchers in this field of interest.
5. Conclusions and Future works: final results of this research and future work suggestions.

2. Background

In this section we review the Mobile Cloud Computing (MCC), Market Basket Analysis (MBA), Association Rule Mining and Location-based services

2.1 Mobile Cloud Computing (MCC)

Mobile cloud computing consists of three modules: Mobile devices, communication network and cloud as a server. The application which is large can be decomposed to smaller ones to process concurrently. This method is called as application partition. Offloading is the process of transferring Mobile application to cloud. This saves the device memory, processing power, and ultimately battery consumption. The classic facilities needed by a Mobile cloud client are, synchronization, push, i.e., updates the notifications sent by the cloud server, offline App automatically handles synchronization and notification, network, database, inter App Bus; provides low-level coordination between applications.[15]

2.2 Market Basket Analysis (MBA)

Market Basket Analysis is a forming technique based on the theory that if you buy a certain group of items, you are more likely to buy another group of items. MBA uses this information to: (1) Understand why they make certain purchases, (2) Products which are purchased together, and (3) Products which might benefit from the promotion. This system has used the Market Basket Analysis method for analyzing the data. The following techniques are used in the analyzing process.[14]

2.3 Association Rule Mining

MAD-ARM is the Mobile Agent-based distributed data mining architecture. It contains knowledge server which works on the generation of association rule and data coming are from different remote sites. The item sets are always upgrading on remote sites at the immobile agent. [3]

The association rules have been recently recognized as an important tool for knowledge discovery in databases. The problem of discovering association rules was first investigated in pioneering work in [4]. Here we examine a database of records which consist of both customer profile (such as salary and age) and behavior (such as buying decision) information). The association rule problem was originally proposed for the case of binary itemset data.[9] The intuitive implication of the association rule X ~ Y is that a presence of the set of items X in a transaction also indicates a strong possibility of the presence of the itemset Y. The measures used to evaluate the strength of the rule are support and confidence. The *support* of a rule X ~ Y is the fraction of transactions which contain both X and Y. The *confidence* of a rule X ~ Y is the fraction of transactions containing X which also contain Y. A considerable amount of research effort [5, 6], has been devoted to the problem of speeding up the itemset method for finding association rules from very large sets of transaction data. Several variations of this problem have been discussed in [7, 8]. The quantitative association rule problem is a generalization of the problem of discovering association rules on sales transaction data, in which both categorical and quantitative attributes are allowed.[10]

2.4 Location-based services

The Mobile location-based application for any shopping product was designed and developed to find nearby branch and stores in local markets, the association among the products purchased, display association to customer side screen, post order, and it is deployed on the Cloud (See Figure1). The grouping of web map service API and association rule mining using Mobile in the cloud, it is potential to gather large-scale shopping habit of people, with lower data collection cost. The developed application model represents an environment for data analysis, and the proposed Algorithm is chosen dynamically at each handset. This is based on the environment where data stream mining process runs on user's android handset. As the data streams in continuously, possible concept drift is updated. There is a specific central Mobile decision agent

which switches several others stream mining agents. Stream mining agents working on local Mobile phones decides the best possible algorithm to run on the local data.

3. Proposed Architecture

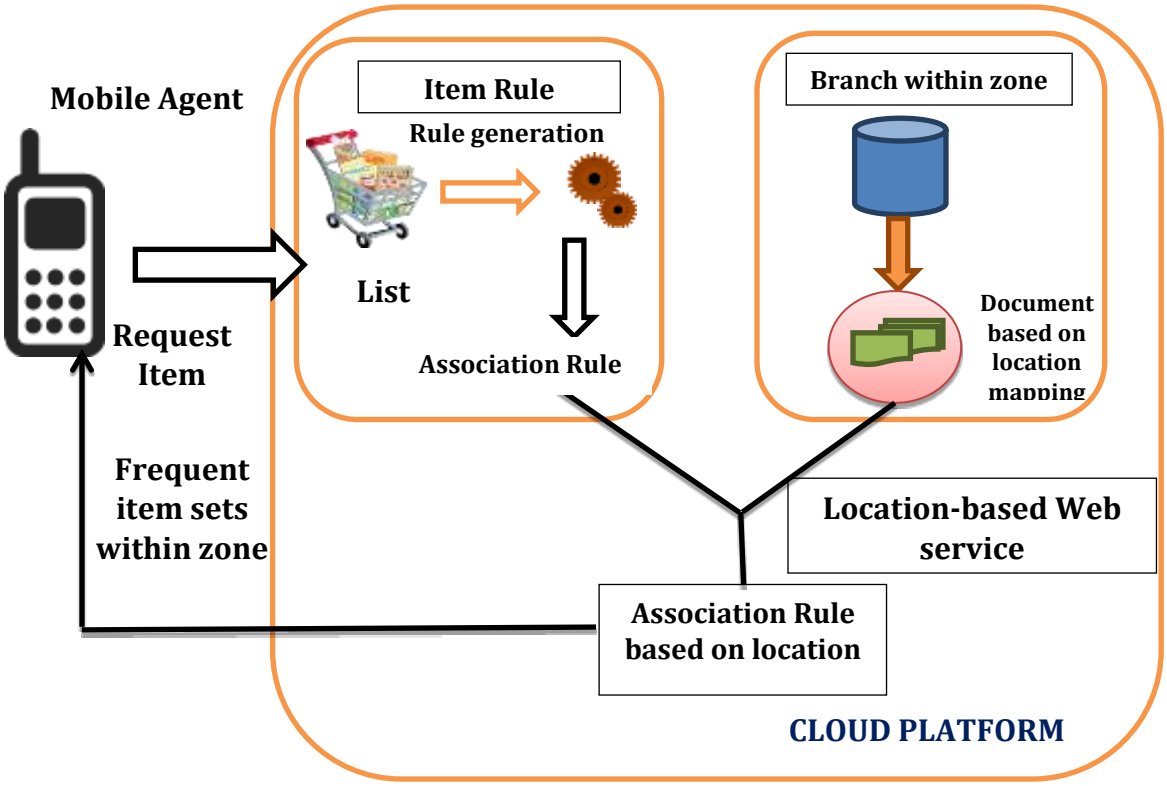


Figure 1 - Proposed Architecture Mobil Agent Association Mining System for Shopping

Any product or application that uses the location data of Mobile subscriber is called as location-based service. Location-based services like GPS uses the latitude and longitude data. Defines a location-based text mining approach to categorize texts into numerous categories based on their geospatial features, with the goals of discovering relationships between documents and area. There are three main modules in this architecture, including geographic data group and reprocessing, mapping forms into corresponding regions, and framing maximizes zones. Data mining and processing are takes place based on zones. The tourism industry has also taken the benefit of location-based services. This application is designed and established using the cloud based platform. It discovers out the location of tourists, where they are directed or looking.

3.1 System Architecture

The architecture of shopping application is divided into three parts as Front end, Middle ware level, and backend.

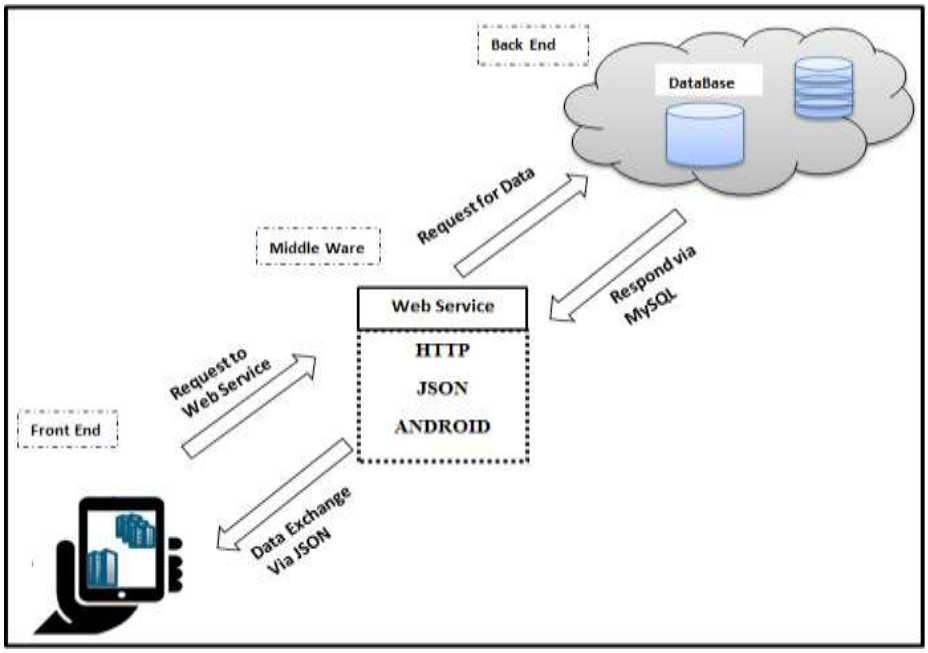


Figure 2: System Architecture. [11]

The Android open source platform is used to design and develop the shopping application. For the end user, in front-end user can able to select particular supermarket branch in the city and see the offers available in the specific showroom at the mall. It also provides offers of local market which are available to users nearby location. The registered retailers can upload and remove their offers and advertisement from this application. When the user searches for specific offers of the showroom in a mall, then the request is sent to middleware level that is to the Web Service. Web service acts like interface for front-end and back-end. The data exchange between front-end and back-end of shopping application happens via middleware level. Android shopping application

sends HTTP Request, and Web Service will send Query to fetch requested data from MySQL database located on the cloud (See Figure 2).

3.2 Software and data Set

In this paper we used data set form supermarket that contains 4700 Records, every record has 23 Items for purchasing case registration. By using and prepare the rules for Apriori diagram and test this data by using Weka application. The Weka is machine learning algorithm for data mining. Weka is efficient and has a user-friendly user interface. It is fully implemented by Java language there for it runs on almost any computing platform. However, it can only run in the local environment.

Apriori Algorithm

- **Product set:** a set of items
- **k-product_set:** a product set which consists of k items

  Frequent product_set (i.e. large product_set): an product_set with sufficient support
- **Lk or Fk:** a set of large (frequent) k-product_sets
- **ck:** a set of candidate k-product_sets
- **Apriori property:** if an item A is joined with item B, Support(A U B) = min(Support(A), Support(B))

In the process of Apriori, the following explanations are needed:

Definition 1: Suppose T={T1, T2, … , Tm}, (m_1) is a set of transactions, Ti= {I1, I2, … , In},(n_1)

is the set of items, and k-product_set = {i1, i2, … , ik},(k_1) is also the set of k items, and k-product_set ⊆I.

Definition 2: Suppose_ (product_set), is the support count of product_set or the frequency of occurrence of a product_set in transactions.

Definition 3: Suppose Ck is the candidate product_set of size k, and Lk is the frequent product_set of size k.

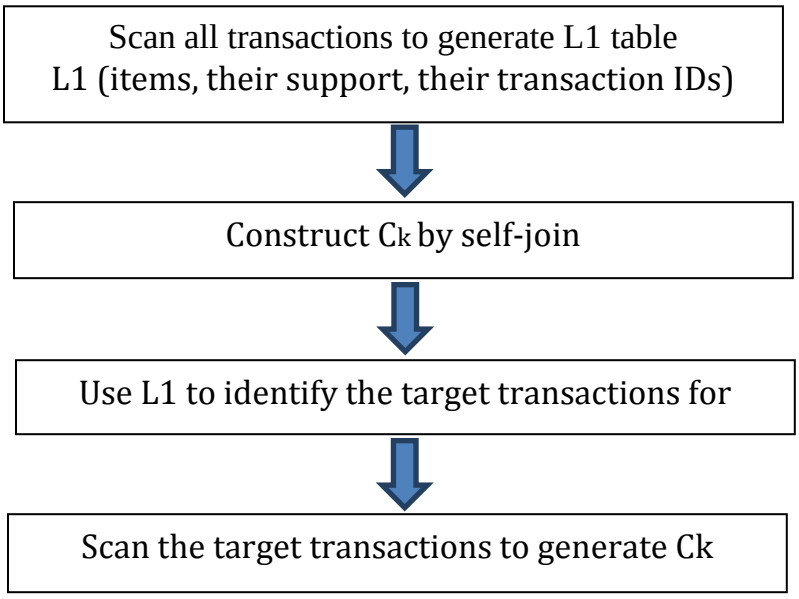


Figure 3: Steps for Ck generation

In the proposed approach, scan all transactions to create L1 which contains the products, their support count and Transaction ID where the products are found. And then use L1 later as a helper to generate L2, L3 ... Lk. When to create C2, make a self-join L1 * L1 to construct two product_set C (x, y), where x and y are the products of C2. Before scanning each transaction records to count the support count of every candidate, use L1 to get the transaction IDs of the least support count between x and y, and thus scan for C2 only in these specific transactions. The same thing for C3, construct three product_set C (x, y, z), where x, y, and z are the products of C3 and use L1 to get the transaction IDs of the least support count between x, y, and z, then scan for C3 only in these specific transactions and repeat these steps until no new frequent product_sets are identified. The whole process is shown in (Figure 3).

4. Related Work

- In [11] this paper designed and established a location-based mobile shopping application for malls and local markets for android platform. This application shows nearby local markets and mall's stores that are registered to the application. The main objective of marketing is achieved at a very low cost in comparison of advertisements, announcements, ground level marketing etc.
- In [12] this study has represented a recommendation engine by using association rules. The system had been tested by existing data in terms of the accuracy and the coverage. Best results have determined for 25 days data with 87.74% coverage and 16.43% accuracy. Processing time was 318 minutes for those data.
- In [13] this paper, we discussed the problem of online mining of profile association rules. Such rules may be very useful in developing relationships between consumer profiles and behavioral information. We discussed how to use multidimensional indexing to generate profile association rules in online fashion.

5. Conclusions and Future works

We designed this architecture after survey of the research on the use of mobile in the purchase process using an application on the cloud computing systems and tried to improve this service, also to monitor the influence of the factor of space and distance between the consumer and supermarkets and calculated the distances between the place of contact of the consumer and the nearest branches of supermarkets and the lowest prices in the same Time and therefore there are more opportunities for the consumer. We designed a mobile shopping architecture based on the site for the very large supermarket has many Branches of the Android platform. The main objective of marketing is achieved at a very low cost compared to advertising, and marketing at the regional level and so on. Data is exchanged between different levels of structural design using a web service station that generates a JSON format for data transfer. With Mobile Computing Cloud Computing Mobile processing and storage move to the cloud as a server, helping to save battery consumption and extend performance or speed of implementation. In this paper we designed architecture for the use of mobile in the purchase process of the cloud computing system and monitored the influence of the distance between the mobile and the nearest supermarket branch to the consumer. In future works, we can focus on small organization more precisely. The architecture can feature more options for live stations for small household business. A broad

market survey is needed to be done to provide marketing and selling for such developers and products.

REFERENCES

[1] Emil Almrot, Sebastian Andersson. A study of the advantages & disadvantages of mobile cloud computing versus native environment. Bachelor Thesis in Software Engineering May 2013.

[2] Raubal, M. and Rinner, C., 2004. Multi-Criteria Decision Analysis for Location Based Services. In S.A. Brandt (ed.) Proceedings of the 12th International Conference on Geoinformatics, pp. 47-53.

[3] A.Saleem Raja, E.George Dharma Prakash Raj, "MAD-ARM: Mobile Agent based Distributed Association Rule Mining", International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 09 – 11, 2013, Coimbatore, INDIA, DOI 10.1109/ICCCI.2013.6466112.

[4] Agrawal R., Imielinski T., and Swami A. 1993. Mining association rules between sets of items in very large databases. Proceedings of the A CM SIGMOD Conference on Management of data, pages 207-216, Washington D. C.

[5] Agrawal R., and Srikant R. 1994. Fast Algorithms for Mining Association Rules in Large Databases. Proceedings of the 20th International Conference on Very Large Data Bases, pages 478-499.

[6] Savasere A., Omiecinski E., and Navathe S. 1995. An Efficient Algorithm for Mining Association Rules in Large Data Bases. Proceedings of the 21st International Conference on Very Large Data Bases. Zurich, Switzerland, pages 432-444.

[7] Hun J. and Fu Y. 1995. Discovery of Multiple-Level Association Rules from Large Databases. Proceedings of the 21st International Conference on Very Large Data Bases. Zurich, Switzerland, pages 420-431.

[8] Agrawal, R. and Srikant, R. Mining Generalized Association Rules. Proceedings of the 21st VLDB Conference. 1995

[9] Agrawal R., Imielinski T., and Swami A. 1993. Mining association rules between sets of items in very large databases. Proceedings of the A CM SIGMOD Conference on Management of data, pages 207-216, Washington D. C.

[10] Srikant R., and Agrawal R. Mining quantitative association rules in large relational tables. Proceedings of the 1996 ACM SIGMOD Conference on Management of Data. Montreal, Canada.

[11] Shivani Rajput*, Rajeshree Pagar, Laxman Erande, Yogesh Alai. One Click Android Application for Shopping Based on Cloud. Proceedings of the 2015 IJSRSET | Volume 1 | Issue 5 | Print ISSN: 2395-1990 | Online ISSN: 2394-4099.

[12] Ozgur Cakir, Murat Efe Aras. A recommendation engine by using association rules Procedia - Social and Behavioral Sciences 62 (2012) 452 – 456.

[13] Charu C. Aggarwal, Zheng Sun, Philip S. Yu. Online generation of profile association rules - T.J. Watson Research Center Yorktown Heights, NY 10598 charu@watson.ibm.com, Duke University, Durham, NC-27706 sunz@cs.duke.edu, T.J. Watson Research Centerm, Yorktown Heights, NY 10598 psyu@watson.ibm.com.

[14] Manpreet Kaur, Shivani Kang. Market Basket Analysis: Identify the changing trends of market data using association rule mining, Procedia Computer Science 85 (2016) 78 – 85.

[15] S M Shamim, Angona Sarker, Ali Newaz Bahar, Md. Atiqur Rahman. A Review on Mobile Cloud Computing. International Journal of Computer Applications (0975 – 8887), Volume 113 – No. 16, March 2015.

Mining IoT Data Streams

Zaki. A. M.[1]

Abstract

Internet of Things "IoT" is the natural evolution of wireless sensor networks, mobile computing, and cloud computing. IoT is the result of trying to benefit from networking advances, hardware advances by connecting every possible object to the internet. The objective of IoT is to make every connected object intelligent enough to take decision to help people in their live. In this paper the concept of mining IoT data streams is reviewed. The mentioned data streams are the streams of data that are captured by or from IoT's objects. This research gives a short background on IoT and state a set of research questions to answer. Then, a discussion to answer research question is conducted.

Keywords

IoT, Data Streams, Edge Analytics, Edge Mining

1. Introduction

In the last decade and due to advances in networking and virtualization a new term called cloud computing has arises. In cloud computing, processing and storage moved to the cloud and I/O, lite processing and lite storage remains on the client. Cloud computing goal was to provide a powerful and scalable services, platforms and infrastructures to satisfy business needs on demand [Mell, 2011].

On the other hand, there were advances in mobile computing. Mobile devices – including sensors, mobile phones, tablets, and laptops – have become more powerful in processing, memory, storage, networking and battery. Additionally, huge number of these mobile devices are produced and connected to Internet. Not only mobile devices which is connected to Internet. There are other products that are now contain an embedded computer and connected to internet such as cars, refrigerators, watches, TVs, air conditioner, satellite receivers, traffic lights, surveillance

[1] a.m.zaki@sadataacademy.edu.eg
Information Systems and Technology Department, Institute of Statistical Studies and research, Cairo University

cameras and weather sensors. All mentioned devices generate streams of data with different types. While these objects are connected to internet a new term coined which is Internet of Things "IoT". Traditionally generated data is transferred to backend systems to be cleansed, stored, and analyzed.

Machine learning algorithms are used to make these things "Objects" intelligent. A new approach to handle these data is to bring intelligence on the edge of the network. The new approach called edge analytics. In edge analytics approach, data streams are processed during the journey to the cloud [Gaura, et al., 2013] [Simoens, 2015] [HP, 2017]

The objective of this research is to provide a complete picture of current literature in mining IoT data streams at the edge of networks. This research tries to answer a set of questions stated in section 2. In section 3, IoT concept is discussed. Section 4 shows big data analytics approach and its drawbacks. Section 5 introduces the concepts of edge computing, edge analytics. Section 5 also discusses edge computing motivations, challenges, opportunities and applications. Section 6 reviewed data stream mining concepts and shows the algorithms and models of data mining used with IoT. Finally, in section 7, Open issues of IoT is discussed.

2. Research Questions

Research questions to be addressed are:

Table 1 Research Questions

RQ#	Description
RQ1:	What are the existed approaches to mine data streams from IoT devices?
RQ2:	What are the reasons leading to the need to make analysis on the edge device?
RQ3:	Is there any algorithm that is already adapted to work on edge devices?
RQ4:	What are the benefits that returns by mining data streams on edge devices?

These questions could be expanded to be as follow:

RQ[1]. What are the existed approaches to mine data streams from IoT devices?

- RQ[1.1] What are IoT Devices?
- RQ[1.2] What are IoT devices characteristics?
- RQ[1.3] How much Data IoT Devices collect?
- RQ[1.4] How to get benefit from this data?
- RQ[1.5] What are the different approaches to handle these data?

RQ[2]. What are the reasons leading to the need to make analysis on the edge device?

RQ[2.1] How traditional approach for handling IoT data Streams works?
RQ[2.2] What are the drawbacks of the traditional approach?

RQ[3]. Is there any algorithm that is already adapted to work on edge devices?

RQ[3.1] What are the existed machine learning algorithms for mining data streams?
RQ[3.2] Which algorithm could be adapted to work on edge devices?
RQ[3.3] What are algorithms that are already adapted to work on edge devices?

RQ[4]. What are the benefits that returns by mining data streams on edge devices?

RQ[4.1] What are the different applications of Edge Analytics?
RQ[4.2] What are the benefits from these kinds of applications?
RQ[4.3] What are the side effects of these applications on society?

3. Internet of Things

In this section, researcher tries to identify what Internet of Things "IoT" is and what IoT devices are. A lot of researches try to answer these questions. IoT could be defined as "The worldwide network of interconnected objects uniquely addressable based on standard communication protocols" which is RFID group definition according to [Gubbi, et al., 2013]. This definition is a network based definition which focuses on communication protocols and networking. It ignores the nature of connected object. Another view of IoT is an intersection between Internet, Thing and semantic oriented visions [Atzori, et al., 2010]. Figure 1 represents the previous definition as it considers the nature of object and semantics of environment and communication standards [Atzori, et al., 2010].

Gubbi give more general definition for IoT. They define IoT as "Interconnection of sensing and actuating devices providing the ability to share information across platforms through a unified framework, developing a common operating picture for enabling innovative applications. This is achieved by seamless ubiquitous sensing, data analytics and information representation with Cloud computing as the unifying framework" [Gubbi, et al., 2013] . From the previous three definitions, IoT is to connect objects "Things" together with a cloud based backend systems over the internet, these things can sense, generates and capture data and then transmit that data to the cloud to be analyzed, visualized and stored. Sometimes these objects have an actuator to do tasks. IEEE defines IoT as "Internet of Things envisions a self-configuring, adaptive, complex network that interconnects 'things' to the Internet through the use of standard communication protocols. The

interconnected things have physical or virtual representation in the digital world, sensing/actuation capability, a programmability feature and are uniquely identifiable. The representation contains information including the thing's identity, status, location or any other business, social or privately relevant information. The things offer services, with or without human intervention, through the exploitation of unique identification, data capture and communication, and actuation capability. The service is exploited through the use of intelligent interfaces and is made available anywhere, anytime, and for anything taking security into consideration." [Initiative, 2015] .In the previous definition IEEE was trying to define IoT with covering all perspectives and points of view.

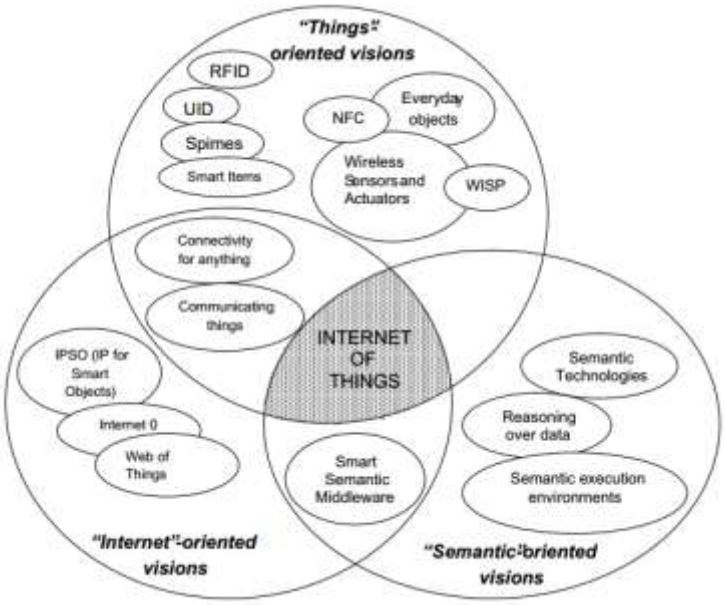


Figure 1 'Internet of Things" paradigm as a result of the convergence of different visions, Source: [Atzori, et al., 2010]

The next question is what type of data that could be captured are and how much this data is and how to deal with this data. Types of data that could be captured are text in all formats, images and videos. Traditionally, all data captured and send to cloud over network. When the cloud receives those data, backend systems starts to clean and store it in form that facilitates analysis, visualization process.

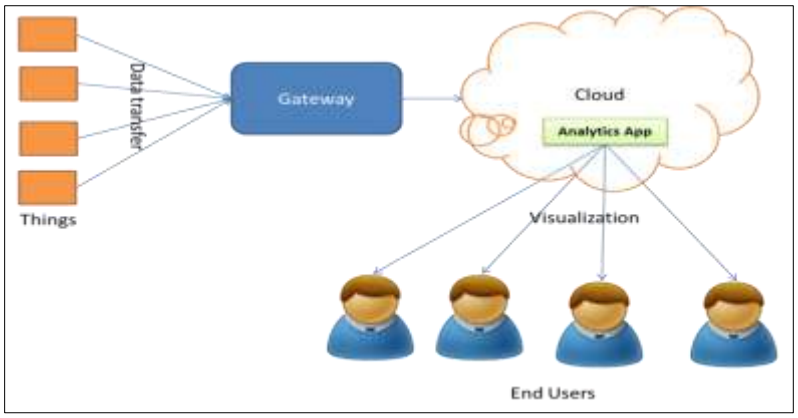


Figure 2 Simple illustration of Big Data Analytics

Actually, data type is not the problem. The major problem is data size, there is a forecasting tells that by the year 2020, there will be 26 billion connected devices and data size will reach to 40 zettabytes as illustrated in Figure 3 [Siemens, 2017].

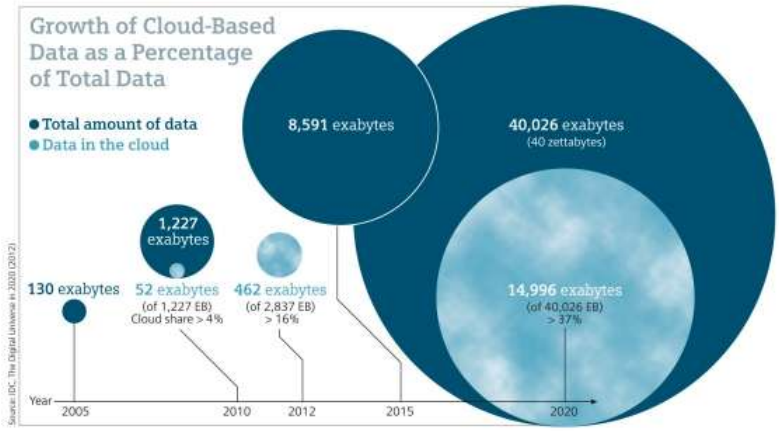


Figure 3 Growth of Cloud Based data as a Percentage of total data source [Siemens, 2017]

With this massive amount of data, the question arises, how to handle this amount of data. The next question, dose it worth to store these huge amounts of data. These problems will be discussed in the next section.

4. Big Data Analytics approach and drawbacks

As discussed earlier, the existed approach of big data assume that data is collected from sensors and other things, and then transferred to backend system on the cloud. These systems are responsible for preprocess, store, analyze and visualize analysis results to users.

This approach has its own drawbacks. These drawbacks are seven according to HP [HP, 2017]: latency, bandwidth, cost, threats, duplication, corruption and compliance. These are the drawbacks of existing approach and reasons for moving towards edge analytics. Table 2 summarizes these seven drawbacks. From researcher point of view these drawbacks may be solved by moving analysis into edge. But other problems may occur.

Table 2 : drawbacks of Existing approach summary

Drawback	Description
Latency	Network latency may cause a disaster in critical real time application. Where in existing approach it may take minutes before response.
Bandwidth	With huge amount of data, it consumes bandwidth to upload data on to the cloud which will also affect network latency.
Cost	To send huge amount of data to cloud, it needs high bandwidth which costs a lot of money. In some cases, it is not needed to move all data from sensors into backend.
Threats	Security Threats like while sending data over network (security attacks)
Duplication	Data could be duplicated while sending to the cloud
Corruption	During data transmission, data may be get corrupted due to any network issue.
Compliance	In some critical real time application, it is not allowed to move data outside country boundary. If the cloud is outside country this will cause a problem.

5. Edge Analytics

Edge analytics refers to the processing on the edge device at the edge of the network. In simple words, processing and analysis is done before sending anything to the cloud. It could be done on the edge "IoT" device or on the network gateway. According to [Varghese, et al., 2009], there are two layers of edge computing devices. The first is edge device which includes "Things". The second layer is edge nodes which include networking devices such as routers, switches and base stations.

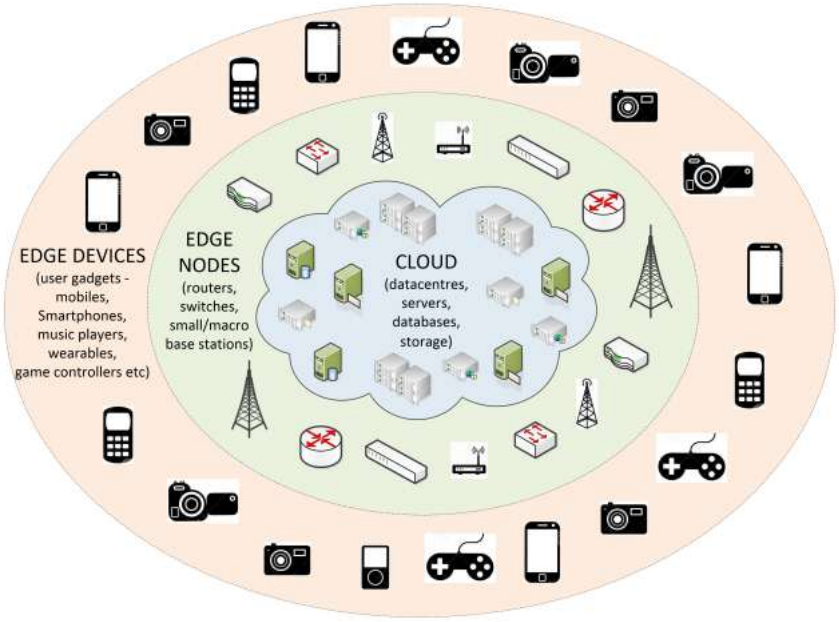


Figure 4 Edge Devices, Edge nodes and Cloud

source [Varghese, et al., 2009]

Edge analytics provide a solution for the previous drawbacks of big data model. But this also faces challenges. One of these challenges is that IoT devices are limited or constrained in resources compared with resources on the cloud. IoT devices may face a problem in power consumption and battery size. If analysis happened on network gateway, this will solve the power problem. But lead to another problem which is network gateways is a specific purpose computer system. Which means it needs to be modified or replaced with another system capable to do analysis. Another challenge is to adapt algorithms to run on such devices. Challenges and opportunities are discussed in more details in next subsection.

5.1. Edge computing motivations, challenges and Opportunities

Varghese proposes summary of motivation, challenges and opportunities in edge computing in Figure 5 [Varghese, et al., 2009]. According to [Varghese, et al., 2009] , motivations of edge computing are decentralizing cloud, enhancement in computational resource of client devices, huge energy consumption at the cloud, data explosion and smart computation techniques. Each of these motivations is discussed in

Table **3**.

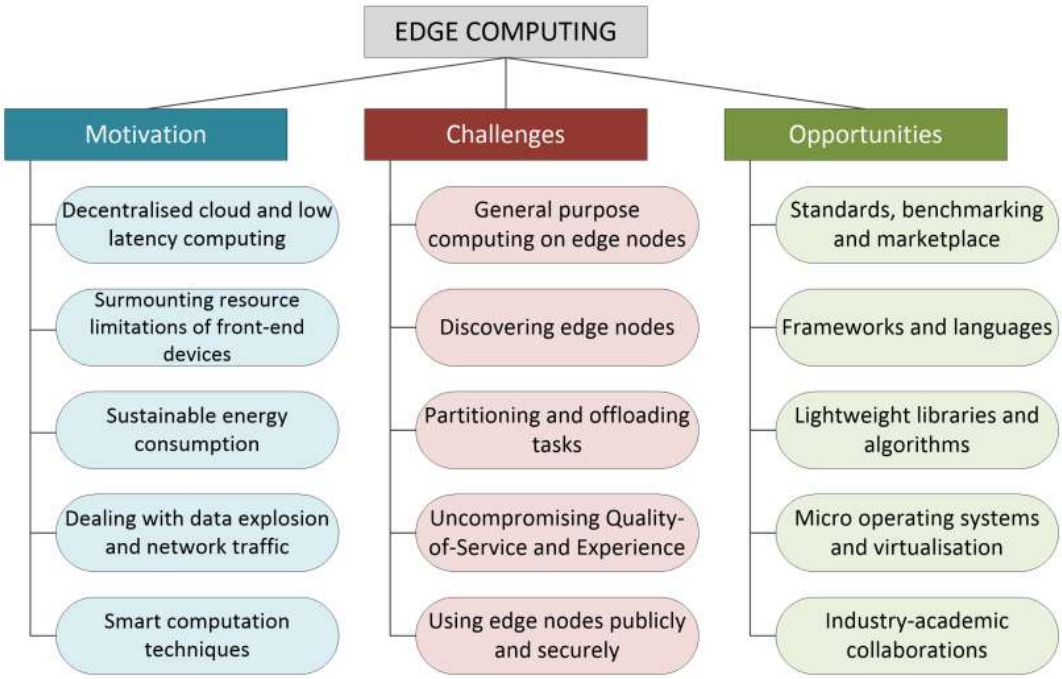


Figure 5 Edge Computing Motivation, Challenge and Opportunities

Source: [Varghese, et al., 2009]

Table 3: Edge Computing Motivation

Motivation	Description
Decentralized cloud	In real time applications, using central cloud to process data stream dose not fulfill time requirements. So, moving the processing to edge node or edge device satisfy time constrains. Decentralization is useful in such cases.
Enhanced Resources in Client Devices	Recently, edge devices could be considered resource limited if it compared to servers or workstations. But it is not resource limited to do filtrations and analytics.
Sustainable energy consumption	Data centers at Cloud consume huge amounts of energy and this amount is increasing. By using edge devices and nodes in some tasks, this leads to minimizing energy consumption at the cloud.
Data Explosion and network traffic	As mentioned earlier, data sizes are increasing dramatically due to the increase in number of connected devices. This affects network bandwidth while transferring data to the cloud. By moving processing on the edge, this help in minimizing data to transferred to the cloud.

Motivation	Description
Smart computation techniques	Using computations techniques to distribute the application using the idea of pipelining the processing in horizontal fashion in which the processing is starting at edge device, then edge node and finally at the cloud. Another technique is to use computational offloading in which the cloud offloads computational tasks to edge devices and node to be processed.

On the other hand of these motivations there are challenges that face edge computing. These challenges according to [Varghese, et al., 2009] are general purposes computing on edge nodes, edge node discovering, task partitioning and computational offloading, quality of services issues and edge nodes security concerns. Table 4 describes a summary of these challenges.

Table 4: Edge Computing Challenges

Challenge	Description
General Purpose computing on edge node	Edge nodes like routers, switches and base stations are not general purposes computers. According to that it is not designed to do any processing except the one it designed for.
Discovering edge node	Automatic resources discovery services for edge nodes are not existed. All applied methods in cloud environment do not apply on edge nodes or devices.
Task Partitioning and Computational offloading	Task partitioning or task offloading refers to the idea of distributing application between edge node or device and cloud, the major challenge here is how to determine which portion of task will be executed and where .
Quality Of Services "QoS"	This challenge is to use edge node or device efficiently to meet user expectations in service level without overloading edge node or device.

Finally, Varghese addresses a set of opportunities that could be done in academic researches. These opportunities are first creating benchmarks, standards, frameworks and toolkits for edge computing. Another opportunity is to create lightweight libraries and algorithms that fit edge

nodes and devices. Additional research area is micro operating systems, mobile containers for edge nodes and devices. [Varghese, et al., 2009]

5.2. Edge Analytics Applications

There are a lot of edge analytics on IoT Applications. A sample application of edge analytics on IoT is Surveillance cameras. According to [Simoens, 2015], surveillance cameras are common in police cars and also available for commercial uses. In [Simoens, 2015], they propose an architecture called GigaSight for video streaming analytics purposes. Another application of edge analytics is health care. Rahmani et al, propose architecture for an IoT-based health monitoring that could be deployed inside hospitals or homes. It depends on health monitoring sensors and smart e-health gateway. These edge devices and node do the analysis and call for help on time [Rahmani, et al., 2015] . Another example is smart city which introduced in [Bélissent, 2010] . According to Bélissent, IoT could be used in areas such as transportation, healthcare, education, public safety and security, building management, city administration and waste management. By adopting IoT in those areas, smart city will be existed. Bélissent also provide a comparison between new cities, existed cities and non-cities from the point of obstacles to apply smart city.

6. Data Stream mining

Data stream mining is concerned with mining streams of real time data generated from sensors in Wireless Sensor Networks "WSN". Data Stream mining is one of two categories of data streams processing. According to [Gaber, et al., 2010] data streams processing has two categories which are data stream management and data stream processing. Data stream management is preparing data stream to other processing by querying and summarizing it. Whereas data stream mining is to apply data mining techniques on this stream of data. According to [Ullman, et al., 2014] , data stream management system has a main component for dealing with streams which is stream processor. Stream processor handles ad-hoc queries on stream input and store data which is needed and other data dose not stored. Figure 6 shows the idea of data streams management system.

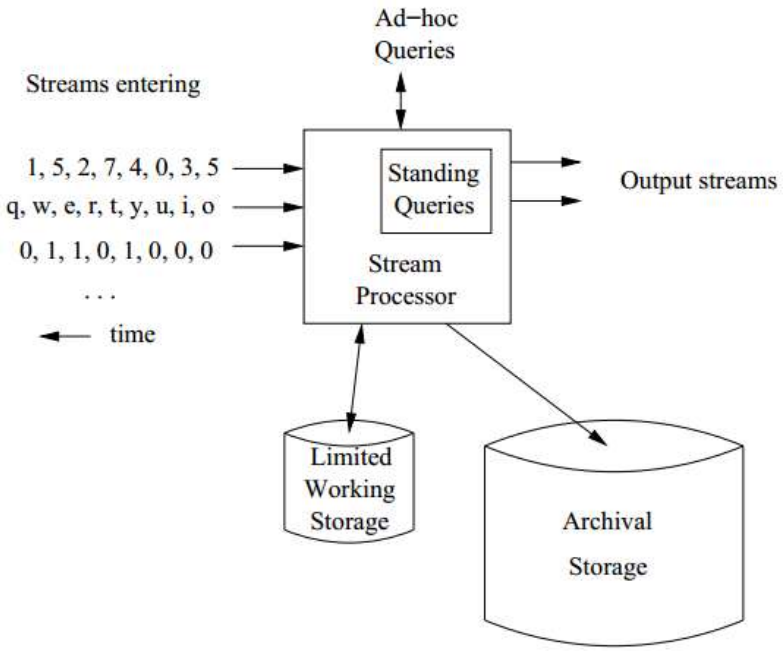


Figure 6 - Data Stream Management System
Source **[Ullman, et al., 2014]**

6.1. Data Mining for IoT

According to [Shen, et al., 2010] , there are four proposed data mining models for IoT. These four models are multi-layer model , distributed model , grid based model and multi-technology based model. In multi-layer model there four layers: data collection layer, data management layer, event processing layer and data mining service layer. Data collection layer is responsible for collecting objects data using RFID readers, GPS and any other adopted sensor. The next layer is data management layer which responsible for managing collected data using centralized or distributed database or warehouse. Then collected data is cleaned and processed to produce analytics. The third layer is event processing layer which is responsible for events filtering and events detection. This layer filters all unrequired events and keeps events that users are interested in. Finally, the last layer is data mining service layer. Data mining service layer is responsible for build learning models using data mining techniques to extract knowledge from objects data and events data [Shen, et al., 2010].

On the other hand, distributed data mining model tries to cope with IoT data characteristics, hardware limitations and security requirements

[Shen, et al., 2010]. According to [Shen, et al., 2010], distributed data mining model preprocess data locally at node and send to central backend system only required and necessary data for processing. The third model according to [Shen, et al., 2010] is grid based data mining model. In this model, IoT devices are considered as computing resources for a grid. This grid is used later for data mining tasks. The last model that was proposed by [Shen, et al., 2010] was based on mixing networking technologies, sensing technologies and mining algorithms then provide "Intelligence" to applications.

Additionally, [Shen, et al., 2010] discusses challenges of mining IoT data. These challenges were about data collection issues such as data sizes, data transmission and energy utilization. Another challenge is about what model to adopt in data centralized processing or distributed processing. Also [Shen, et al., 2010] refers to the challenge of studying data mining algorithms that are suitable for IoT.

In 2013, a paper entitled "Edge Mining the Internet of Things" define a term "Edge Mining" as "Processing of sensory data near or at the point which it is sensed, in order to convert it from a row signal to contextually relevant information" [Gaura, et al., 2013] . This paper introduces four mining algorithms that used to reduce network traffic and reduce energy consumption. These algorithms are "General Spanish Inquisition Protocol" G-SIP [Gaura, et al., 2013] , "Linear Spanish Inquisition Protocol" L-SIP [Gaura, et al., 2013] , "ClassAct" [Gaura, et al., 2013] and "Bare Necessities" [Gaura, et al., 2013] BN. Figure 7 summarize edge mining process according to [Gaura, et al., 2013]

G-SIP and L-SIP are based on SIP. The basic idea of SIP is to send unexpected information to control center or to base station [Gaura, et al., 2013]. ClassAct is a classifier for human posture [Gaura, et al., 2013] . BN is an algorithm that calculates time spent in specific state [Gaura, et al., 2013].

In 2014, a survey paper about data mining for internet of things. This paper reviews motivation and problems then reviewed solution and results. Finally, it gives future directions and trends in this area. This paper was written by [Tsai, et al., 2014]

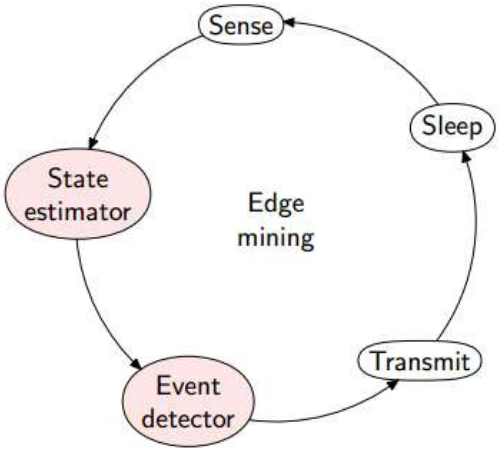


Figure 7: Edge Mining Process at node

Source: **[Gaura, et al., 2013]**

In [Tsai, et al., 2014] , researchers start with describing data captured from IoT and what this data about and data sizes. Tsai and the others refer to possible solutions of handling big data generated from IoT. Additionally, they refer to the benefits gained from analyzing IoT data. They also provide architecture for knowledge discovery and IoT. Figure 8 illustrates Tsai architecture [Tsai, et al., 2014]. This architecture starts with collecting data and preprocesses it then extracts knowledge then update the IoT device with that knowledge.

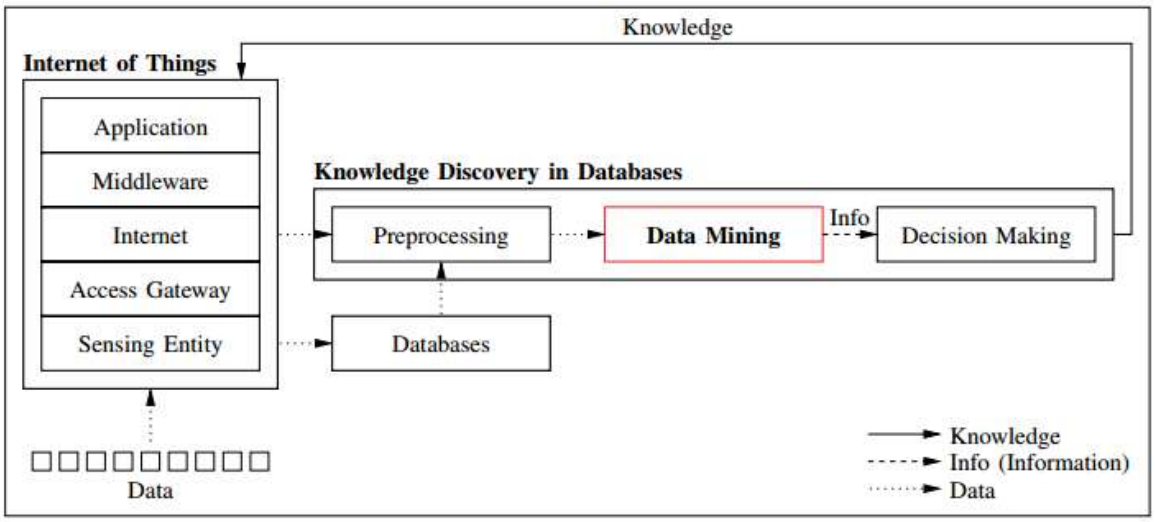


Figure 8: Architecture of IoT with KDD
source **[Tsai, et al., 2014]**

In [Tsai, et al., 2014], there are three considerations for choosing suitable data mining technique for IoT. These Considerations are objective, data and mining algorithm. Objective is about problem assumptions, limitations and measurement way. Data consideration is about data size, distribution and type. The last consideration is algorithm. The challenge is how to determine data mining algorithm. Additional challenge about algorithm, will the problem require to develop new algorithm or just adopt old one [Tsai, et al., 2014].

Tsai present a framework for data mining, called unified data mining framework. This framework is used to explain all mining algorithm that covered in their research. This framework is based on scan, construct and update operation. Scan input data set, then construct rules as output then update these rules. The previous operations continue until met a termination criterion [Tsai, et al., 2014]. Figure 9 shows the unified data mining framework presented by Tsai.

Algorithm 1 Unified Data Mining Framework

```
1  Input data D
2  Initialize candidate solutions r
3  While the termination criterion is not met
4      d = Scan(D)  [Optional]                    S
5      v = Construct(d, r, o)                     C
6      r = Update(v)                              U
7  End
8  Output rules r
```

Figure 9 Unified data mining framework
source **[Tsai, et al., 2014]**

According to Tsai, data mining techniques could be used for IoT infrastructure or services [Tsai, et al., 2014]. They discuss data mining techniques in terms of the unified data mining framework. Their study investigates each data mining technique for IoT from two perspectives. First perspective is infrastructure perspective (i.e. the design consideration of algorithm to fit in IoT). The second prospective is which is the service algorithm will be used in (i.e. used in which application). The researcher illustrates this taxonomy in Figure 10.

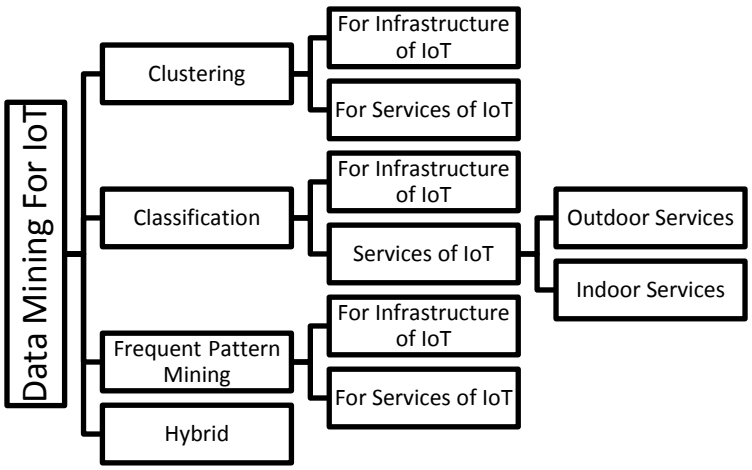


Figure 10 data mining techniques for IoT

Moreover, [Tsai, et al., 2014] provide a comparison of mining technologies in terms of goals and data sources and in reference that study each mining technique. Figure 11 is adapted from [Tsai, et al., 2014] to

show only goal and data source for each mining technique.

Mining Algorithm	Goal	Data Source
Clustering	Network performance enhancement	wireless sensor
	Inhabitant action prediction	X10 lamp and home appliances
	Provisioning of the needed services	Raw location tracking data
	Housekeeping	Vacuum sensor
	Managing the plant zones	GPS and sensor for agriculture
	Relationships in a social network	RFID, smart phone, PDA, and so on
Classification	Device recognition	RFID
	Traffic event detection	GPS, smart phone, and vehicle sensor
	Parking lot management	Passive infrared sensor
	Inhabitant action prediction	RFID, sensor, video camera, microphone, wearable kinematic sensor, and so on
	Inhabitant action prediction	Video camera
	Inhabitant action prediction	microphone
	physiology signal analysis	wireless ECG sensor
Frequent Pattern	RFID tag management	RFID
	Spatial colocation pattern analysis	GPS and sensor
	Purchase behavior analysis	RFID and sensor
	Inhabitant action prediction	RFID and sensor
Hybrid	Inhabitant action prediction	RFID and sensor

Figure 11 Mining Techniques Comparison

Source: **[Tsai, et al., 2014]**

7. IoT Open Issues

Finally, Tsai et al refer to open issues of IoT and data mining. These open issues are from four perspectives: Infrastructure, Data, and algorithm and security perspectives [Tsai, et al., 2014] . Table 5 tries to summaries these open issues from the different perspectives.

Table 5 open issues in IoT

Perspective	Open Issues
Infrastructure	• Decentralization • Heterogeneity • Low power • Small memory, limited computation power
Data	• Different data standards, types and resources • Sizes of Data and preserving of data
Algorithm	• Dynamic nature of sensor status affects algorithms • Redesign of algorithm to fit in application. • Computation load and services balance.

Perspective	Open Issues
Security	• Privacy concern while using video cameras • Sensitive information of IoT end User (e.g. health info)

8. Conclusion

In this paper, a systematic literature review is conducted on mining IoT data Streams. The paper discusses the term IoT, edge computing, data stream mining and data mining for IoT. The paper starts with reviewing drawbacks of big data approach for handling IoT data. Then it reviews challenges and opportunities of edge computing. Additionally, Applications of edge analytics is mentioned. Later, concept of data stream mining is reviewed. After that data mining for IoT is discussed. At the end, it could be notices that a lot of challenges still need research effort. This Research area is still open and there are a lot of opportunities. It is a hot topic for research.

References

[1] Atzori Luigi, Iera Antonio and Morabito Giacomo The Internet of Things: A survey [Journal] // Computer Networks. - 2010. - Vol. 54. - pp. 2787-2805. - ISSN: 1389-1286.

[2] Bélissent Jennifer Getting clever about smart cities: new opportunities require new business models [Journal] // Forrester Research, inc. - 2010. - p. 33.

[3] Gaber Mohamed Medhat, Zaslavsky Arkady and Krishnaswamy Shonali Data Stream Mining [Book Section] // Data Mining and Knowledge Discovery Handbook / book auth. Ratanamahatana Chotirat AnnLin, Jessica Gunopulos, Dimitrios Keogh, Eamonn Vlachos, Michail Das, Gautam. - 2010.

[4] Gaura Elena I [et al.] Edge mining the Internet of Things [Journal] // IEEE Sensors Journal. - 2013. - pp. 3816-3825.

[5] Gubbi Jayavardhana [et al.] Internet of Things (IoT): A vision, architectural elements, and future directions [Journal] // Future Generation Computer Systems. - 2013. - Vol. 29. - pp. 1645-1660. - ISSN: 0167-739X.

[6] HP IoT analytics at The Edge [Report]. - 2017.

[7] Initiative I. E. E. E. Internet Towards a definition of the Internet of Things (IoT) [Journal]. - 2015. - p. 27. - ISBN: 9781509019410 ISSN: 09758887.

[8] Mell Peter M. and Grance, Timothy SP 800-145. The NIST Definition of Cloud Computing [Report]. - Gaithersburg, MD, United States : National Institute of Standards \& Technology, 2011.

[9] Rahmani Amir Mohammad [et al.] Smart e-Health Gateway: Bringing intelligence to Internet-of-Things based ubiquitous healthcare systems [Journal] // 2015 12th Annual IEEE Consumer Communications and Networking Conference, CCNC 2015. - 2015. - pp. 826-834. - ISBN: 9781479963904 ISSN: 2331-9860.

[10] Shen Bin, Liu Yuan and Wang Xiaoyi Research on Data Mining Models for the Internet of Things [Journal]. - [s.l.] : IEEE, 2010. - 978-1-4244-5555.

[11] Siemens Internet of Things: Facts and Forecasts [Online] // Siemens.com. - Siemens, 2017. - April 2017. - https://www.siemens.com/innovation/en/home/pictures-of-the-future/digitalization-and-software/internet-of-things-facts-and-forecasts.html.

[12] Simoens Pieter Edge Analytics in the Internet of Things [Journal]. - 2015. - pp. 24-31.

[13] Tsai Chun-Wei [et al.] Data Mining for Internet of Things: A Survey [Journal]. - [s.l.] : IEEE COMMUNICATIONS SURVEYS & TUTORIALS, 2014. - 1 : Vol. 16.

[14] Ullman J. D., Leskovec J and Rajaraman A Mining Data Streams [Book Section] // Mining of massive datasets. - 2014.

[15] Varghese Blesson [et al.] Challenges and Opportunities in Edge Computing [Journal]. - 2009. - ISBN: 978-1-5090-5263-9.

Arabic Named Entity Recognition

Mariam Muhammad*

Abstract

Named Entity Recognition (NER) is an important task that is used in recognizing the proper names in the text such as "Person", "Location", and "Organization". This task is very important in the most fields of the natural language processing (NLP) such as "Question Answering", "Machine Learning", and "Information retrieval". There are more research studied the NER for the foreign language e.g. English but little of them that studied it for Arabic language because Arabic language has complicated morphology and that make the NER task difficult. In this research, we tried to know more about the Named Entity Recognition task and its importance in the NLP fields especially when using it with the Arabic language text. The challenges that faced the Arabic NER were studied and some the solutions for each challenge were introduced. The linguistic resources and tools that support the Arabic NER were presented and a link for each was illustrated. The main approaches of the NER were explained and an example for each approach was given. The evaluation metrics that are used for evaluating the Arabic NER also were presented.

Keywords: Arabic NER, NER Challenges, Linguistic Resources, NER Tools, NER approaches.

1. Introduction

In the 1990s, NER was introduced as a subtask of information extraction (IE) task and became important in more of the research studies [Shaalan, 2014].

The Named Entity Recognition (NER) task aims at detecting and classifying the proper names in the text. The proper names can be person, location or organization names [Nadeau & Sekine 2007]. For example, in the sentence "Mohammed and Ahmed work in IBM Company in Egypt"; "**Mohamed**" and "**Ahmed**" will be identified as Person NE, "**IBM**" will be identified as organization NE, and "**Egypt**" will be identified as location NE.

* Institute of Statistical Studies and Research, Cairo University, Egypt eng_maryamadel@yahoo.com

The NER task helps the Natural Language Processing (NLP) applications such as Information Retrieval (IR), Question Answering (QA) and Machine Learning (ML) to enhance their performance [Shaalan & Raza 2009].

There are some challenges that face the NER task when it's used with the Arabic language because of the characteristics of Arabic language such as the rich morphology and syntax. Absence the capitalization and the short vowels, the ambiguity, and the lack of the resources are some of these challenges.

The annotated corpora and the lexical are the main linguistic resources for the Arabic NER. Also, there are some the tools that support the Arabic NER such as: "GATE" and "MADAMRA".

There are three main approaches are used in the NER are: (1) Rule Based NER that depends on the linguistic rules of the language, (2) Machine Learning Based NER that depends on the features of the NE classes in a large training corpus, and (3) Hybrid Based NER that combine between the two previous approaches.

It is necessary to evaluate the performance of the NER system. There are some the evaluation metrics that used to evaluate the performance of the NER system. These metrics are part of NLP conferences such as "MUC", "CONLL", and "ACE".

This research aims to give a good background on the Arabic NER for the interested researchers.

The remainder of the research is organized as follows. Section 2 discusses the importance of the Named Entity Recognition Task. Section 3 discusses and analyzes the challenges that face the NER when it is used with Arabic language and the solution for each. Section 4 presents the linguistic resources and the tools that support the Arabic NER. Section 5 explains the main approaches of the Arabic NER and gives an example for each approach. Section 6 reviews the evaluation metrics that are used to evaluate the performance of the Arabic NER. Finally, the conclusion remarks and the future work are presented.

2. The importance of the NER in NLP applications

"Names" play important role in any text for detecting, identifying and extracting them [Mohit, 2014]. So, recognizing the names can improve many applications in NLP. The NLP applications includes: Information Retrieval (IR) systems, Information Extraction (IE), Machine Translation (MT), Question Answering (QA) systems and others.

2.1. Integrated the NER with the information retrieval

The information retrieval (IR) is "the task of identifying and retrieving relevant documents from a set of data according to an input query" [Shaalan, 2014]. According to the studies on the information retrieval, there is a strong relation between the named entity recognition task and the retrieval systems.

NER can be useful for the IR through two possible ways: recognizing the "Named Entities" within the input query and determining the relevant document according to the existence of the recognized "Named Entities" within these documents. For example, the word ("الجزيرة"- Aljazeera) can be recognized as an *organization* Named Entity or a *Location* Named Entity corresponding to the word island. So, the correct recognition leads to retrieving and extracting of the relevant documents.

2.2. Integrated the NER with the Machine Translation

The Machine Translation (MT) is the task of translating a text in a language into another language. Improving the Named Entities translation can help in improving the performance of the MT system [Kaddoura, 2010]. The translation of some Arabic person names to Latin languages faces the ambiguity problem; because the Arabic person name can be found as regular words in the language that isn't a named entity type. For example, the word "سعيد" (Saeed) can be used in Arabic text as a noun (non-NEs) that means "Happy", and also as a Person name (Person NEs). So, the translation of the following phrase " محمد سعيد" can be "Mohamed Saeed" or "Mohamed Happy".

2.3. Integrated the NER with the Question Answering

The Question Answering (QA) task is related to Information Retrieval field where the questions are taken as input and the QA system returns brief answers. For improving the retrieved data in the QA system, the NER task is used; that is through identifying the relevant documents and then extracting the correct answers from candidate passages. For example, the words "الشرق الأوسط" may be classified as *Organization* NE or *Location* NE according to the context.

3. Challenges of the NER in Arabic Language

There are several challenges/problems face the NER task when using it in Arabic Language; this section describes these challenges and studies the possible solution for each.

3.1. Absence of Capitalization Problem

Most Named Entities in the Latin languages like English begin with capital letters such as proper names e.g. ("Ahmed", "Mohammed") and abbreviations e.g. ("ACM", "IBM"). But this feature doesn't exist in the Arabic language because the Arabic language can't support the capitalization. Absence of this feature effects on the Named Entity recognition task [Shaalan, 2014].

The solution of this problem

The dictionary lookup method wouldn't be a suitable solution to face this problem, because some of the words can be used as proper nouns and they also can be used as non-proper noun in the text. For example, the Arabic proper name "عين" can be used in the sentence with different meanings as in Table 1.

Table 1 Example of the absence of the capitalization problem

Word meaning	Word Category	Sentence
Ain	Proper-Noun	عين جالوت
wellspring	Noun	عين الماء
eye	Noun	عين الانسان
delimitate/be delimitate	Verb/passive Verb	عين وزيرا للخارجية

[The source (Saad & Ashour 2010)]

So, the solution of this problem is analyzing the context surrounding the Named Entity.

3.2. Absence of Short Vowels

Arabic text contains the diacritics ً ٌ ٍ َ ُ ِ ّ ْ that "affect the phonetic representation and give different meaning to the same lexical form" [Shaalan, 2014]. But nowadays, Modern Standard Arabic (MSA) is written literally, without considering the addition of diacritics [Alotaibi, 2015]. Absence of the diacritics in the Arabic words leads to the ambiguity because different diacritics represent different meanings. For example, the word "قطر" without any diacritics can be recognized as *country* "location NEs" or as *radius* "Measure NEs".

The solution of this problem

Consider the context of the word in order to predict the correct meaning.

3.3. Complicated morphology

Finding many different patterns for the one Arabic word is from the characteristics of the Arabic language where each word can consist of one or more prefixes, a stem or root, and one or more suffixes in different combinations, that leads to complicated morphology.

Table 2 Arabic Patterns and Roots

Affixes in Arabic	Examples
Prefixes of length 3	وللـ ، وال ، وكال ، بال
Length 2 prefixes	ال ، لل
Length 1 prefixes	ل ، ب ، ف ، س ، و ، ي ، ت ، ن ، ا
Length 3 suffixes	تمل ، همل ، تان ، تين ، كمل
Length 2 suffixes	ون ، ات ، ان ، ين ، تن ، كم ، هن ، نا ، يا ، ها ، تم ، كن ، ني ، وا ، ما ، هم
Length 1 suffixes	ة ، ه ، ي ، ك ، ت ، ا ، ن

[The source (Saad & Ashour 2010)]

The solution of this problem

There are two possible ways to resolve this problem

1. By deleting all the affixes and keeping only the root of the Arabic word. For example, the analysis of "وبالقاهرة" word will result "القاهرة" as a location name after deleting the "و" and the "بـ". Although this way is faster, it leads to a loss of valuable information from the affixes [Benajiba et al., 2007].
2. By separating the suffixes with spaces. For example, the word "وعاصمتها" can be segmented into three parts "ها" "عاصمة" "و". This way is more accurate, because it keeps all affixes and then keeps the meaning.

3.4. Ambiguity Problem

The problem of ambiguity can be between two or more Named Entities because the Named Entity phrases can be formed by different POS such as 'common nouns', 'adjectives' or more complex phrases of more than one token. For example, In the sentence "أحمد أباد فاز بالجائزة": the phrase " **أحمد أباد**" can be recognized as "Person NE" or as a "location NE[1]". Table 3 presented more examples for this problem.

[1] **"Ahmed Abad"** is the largest city and former capital of Gujarat, which is a state in India.

Table 3 Ambiguous Examples

Ambiguous example	English translation	Incorrect	Correct
1.6985 فرنك سويسري	1.6985 Swiss francs	Person	Price
15 رمضان الكريم 2005	15th of Ramadan Al karim 2005	Person	Date
جاسم المتحدة للعقارات والصيانة العامه	Jassim united for real estate and general maintenance	Person	Company
1.5 بليون دولار سنغافورة	1.5 billion Singapore dollars	Location	Price
شركة أرامكو السعودية	Saudi Aramco	Location	Company
راشيل فيكتوريا كيون	Racheal Victoria Queen	Location	Person
اليزابيث الثانية في مساءا	In the evening Elizabeth II	Time	Person
... نقطة تحول في سبتمبر سنة 1954 قدم مارتن...	... a turning point in September 1954 Martin presented ...	Measurement	Date

[The source (Shaalan & Raza, 2009)]

The solution of this problem

Rule based approach can be used for resolving this problem. (Shaalan & Raza, 2009) used the heuristic rules for resolving this problem by "preferring one Named Entity type over the other".

3.5. Transliteration Problem

An NE can be transliterated in many ways. The lack of standardization leads to many shapes of the same word that are spelled differently but with the same meaning. Another reason for this is that "Arabic has more speech sounds than Western European languages, which can ambiguously lead to an NE having more variants" [Shaalan, 2014]. For example, the city of "Washington" could be expressed using four forms such (وشنطن، واشنطن، واشنغطن، واشنجطن)

The solution of this problem

Make a standard form/ a canonical form and normalize each occurrence of the variant to this form; this requires a mechanism (such as string distance calculation) for matching between a name variant and its normalized representation.

3.6. Lack of Resources

There are large annotated corpora as well as Arabic lexicons that can be used for implementing and testing the performance of an Arabic NER system; but most of these available Arabic NER resources are expensive and have limited capacity.

The solution of this problem

Researchers depend on their own corpora, which require human annotation and verification. Few of these corpora have been made freely

and publicly available for research purposes as in [Benajiba et al., 2007] and [Mohit, 2014]; whereas others are available but under license agreements.

4. Arabic Linguistic Resources and Tools supporting Arabic NER task

As it is shown in the previous section, the lack of digital linguistic resources and the tools that support the NER represents a challenge for the NER task especially when it is used with the Arabic language.

This section presents the available linguistic resources and the tools that support the Arabic NER. It detects each of them is open access.

4.1. Arabic Linguistic Resources

There are two types of linguistic resources that are commonly used in NER: (1) Corpora and (2) Lexical resources.

4.1.1. Corpora

The datasets or corpora are used to evaluate and compare the systems. For the NER task, we need "large annotated corpus" where every NE has a type assigned to it.

Some NER corpora are available under paid license agreements, for example, "ACE[2]". And others are freely available such as "ANERcorp[3]" that is a Corpus of more than 150,000 words annotated for the NER task [Benajiba et al., 2007].

Fig. 1 shows a sample of annotated corpora where the words of corpora are collected from different resources and each word are classified into its type: "Loc" means Location NE, "PER" means person NE, "ORG" means organization NE, and "O" means other NE. The letters "B, I,L" means the site of the word in the sentence "Begin", "In the middle" and "Last".

[2] https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications

[3] http://www1.ccls.columbia.edu/~ybenajiba/downloads.html

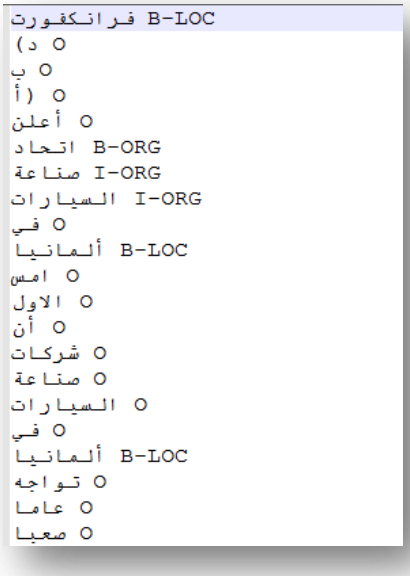


Fig. 1 Sample of annotated Corpora

4.1.2. Lexical Resources (gazetteer)

Gazetteer is another lexical resource which is a collection of predefined lists of typed entities. It can be called as dictionary or wordlist. The contents of a gazetteer should be consistent and belong to only one type of NE. For example, a location gazetteer consists of names of countries, cities, states, and so on [Shaalan & Raza, 2009].

When the researchers found these resources aren't available freely; they built their own gazetteers from different resources such as the Web and from organizations. For example, (Benajiba et al., 2007) built "ANERGazet" that is a collection of 3 Gazetteers:

a) **Locations:** a Gazetteer containing names of countries, cities, states, etc.
b) **People:** a Gazetteer containing names of people recollected manually from different Arabic websites
c) **Organizations:** a Gazetteer containing names of Organizations like companies, football teams, etc.

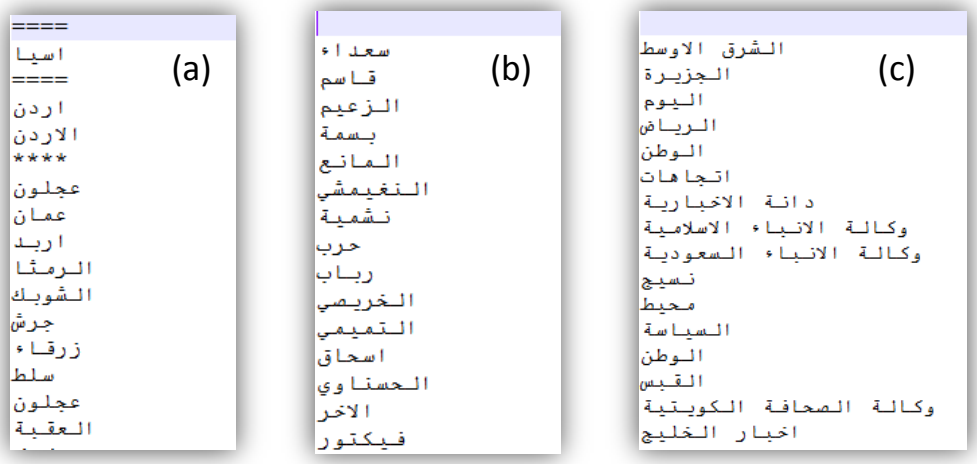


Fig. 2 Sample of the "ANERGazet"
(a) Loc_Gazetteer, (b) Person_Gazetteer and (c) ORG_Gazetteer

Table 4 presents a summary of the important Arabic NER resources that are available to be used.

Table 4 Summary of Arabic NER Resources

Resource name	Resource Type	Availability	Link
"ACE"	Corpora	under paid license agreements	https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications
"ANERcorp"	Corpora	Free	http://www1.ccls.columbia.edu/~ybenajiba/downloads.html
"AQMAR"	Corpora	Free	http://www.cs.cmu.edu/~ark/ArabicNER/
"Fine-grained"	Corpora	Free	https://sourceforge.net/projects/arabic-named-entity-corpora/
"WIKIFANE_Gazet"	Lexical	Free	https://sourceforge.net/projects/arabic-named-entity-gazetteer/?source=directory
"CJK"	Lexical	under paid license agreements	www.cjk.org/cjk/arabic/arabsam.htm
"ANERGazet"	Lexical	Free	http://www1.ccls.columbia.edu/~ybenajiba/downloads.html

4.2. Tools supporting the Arabic NER task

In Arabic language, there is a lack in the NER tools which have more importance for the NLP systems [Kaddoura, 2010].

In this section, some NER tools that have been used in the Arabic NER literature are presented. The tools can be classified into two categories according to their functions: (1) Integrated Development Environments tools and (2) Basic Preprocessing Tools for Arabic.

Table 5 presents a summary of these tools with link for the official site for each.

4.2.1. Integrated Development Environments tools

- **GATE (The General Architecture for Text Engineering):**
  - It is a freely available.
  - It supports these languages (English, French, Italian, German, Arabic, Chinese, Hindi, Romanian, and Cebuano).
  - It can be used to extract basic Arabic entities, such as (date, name, location, organization, and so on).

- **NooJ:**
  - It is a freely available.
  - It supports constructing, testing, and maintaining large lexical resources, and applying morphological analysis for Arabic processing.
  - It can recognize all Unicode encodings.

- **LingPipe:**
  - The free version has limited production capabilities and in order to obtain full production abilities, it must to be upgraded.
  - It is a toolkit for text engineering and processing.
  - It supports POS tagging, spelling correction and NE recognition.

4.2.2. Basic Preprocessing Tools for Arabic

In this section, the Arabic morphological pre-processing tools are presented that are used in the Arabic NER literature, including BAMA, MADA, AMIRA and MADAMIRA toolkit.

- **Buckwalter Arabic Morphological Analyser (BAMA)**
  - It contains over 80,000 words, 38,600 lemmas.
  - It contains three dictionaries (Prefix, Stem, Suffix) and three compatibility tables (Prefix-Stem, StemSuffix, Prefix-Suffix).

- **Morphology Analysis and Disambiguation for Arabic (MADA)**
  - It is a development of BAMA.
  - The TOKAN component allows the user to specify any tokenization scheme that can be generated from disambiguated analyses.
  - The MADA+TOKAN package provides one solution to all of the basic problems in Arabic NLP, including *tokenization*, *diacritization*, stemming, and lemmatization.

- **AMIRA**
  - This is a set of tools including a tokenizer, POS tagger and Base Phrase Chunker.
  - It is used for different NLP applications because of its speed and high performance.

- **MADAMIRA**
  - It combines aspects of two previously used systems for NLP; MADA and AMIRA.
  - It includes several tasks that are useful for NLP processes such as POS tagging, tokenized forms of words, diacritization, lemma stemming, base phrases, and NER.

Table 6 Summary of Arabic NER Tools

Tool	Availability	Link
"GATE"	✓ Free	https://gate.ac.uk/
"NooJ"	✓ Free	http://www.nooj4nlp.net/
"LingPipe"	Free for limited capabilities	http://alias-i.com/lingpipe/
"BAMA"	$ under paid license agreements	https://catalog.ldc.upenn.edu/LDC2004L02
"MADA + TOKAN"	✓ Free	https://lists.cs.columbia.edu/pipermail/mada-users/
"AMIRA"	A demo of the system is available	http://nlp.ldeo.columbia.edu/amira/
"MADAMIRA"	✓ Free	http://www1.cs.columbia.edu/~rambow/software-downloads/MADA_Distribution.html

5. Main Approaches for Arabic NER

According to the research works on the Arabic NER, it is found many approaches for recognizing named entities from text. These approaches have been divided into three categories [Mohit, 2014] as following: (1) Rule Based NER, (2) Machine Learning based NER, and (3) Hybrid NER.

5.1. Rule-Based NER

The Rule-Based is from the early approaches to the NER. It depends on linguistic rules (grammars). It is also called "Linguistic-based" approach [Shaalan, 2014]. There are three main components are used in the Rule-Based systems, as following:

(1) A set of Rules for the named entity extraction task (grammatical rules).
(2) Gazeteers/ Dictionaries that contains different types of named entity classes
(3) Extraction engine that applies the rules to the text.

5.1.1. The advantage and disadvantage of this approach

The advantage of the Rule-based approach is that it needs to high linguistic knowledge to build it.

The problem of this approach is when needing to make any updates or maintenance on the system but the linguists with the required knowledge and background are not available. It requires expensive manual effort and it is time-consuming [Shaalan, 2014].

5.1.2. Example on using this approach

(Shaalan & Raza, 2009) used the Rule-Base approach for developing NER system for Arabic (called NERA system, see Fig. 3). They depended on their own resources, corpora and gazetteers, to train and test the system.

For example, when applying the proposed system in [Shaalan & Raza, 2009] to recognize the NEs in a text, a phrase such "أحمد آباد" can be recognized as a person name or a location. So, it is necessary to create a filter rule to return one correct result. The authors put the following filter rule:

"If a possible match M1 for a location entity reported by the location extractor intersects with a match M2 of a person entity that is also reported by the person extractor, then the match as a location name will be discarded"

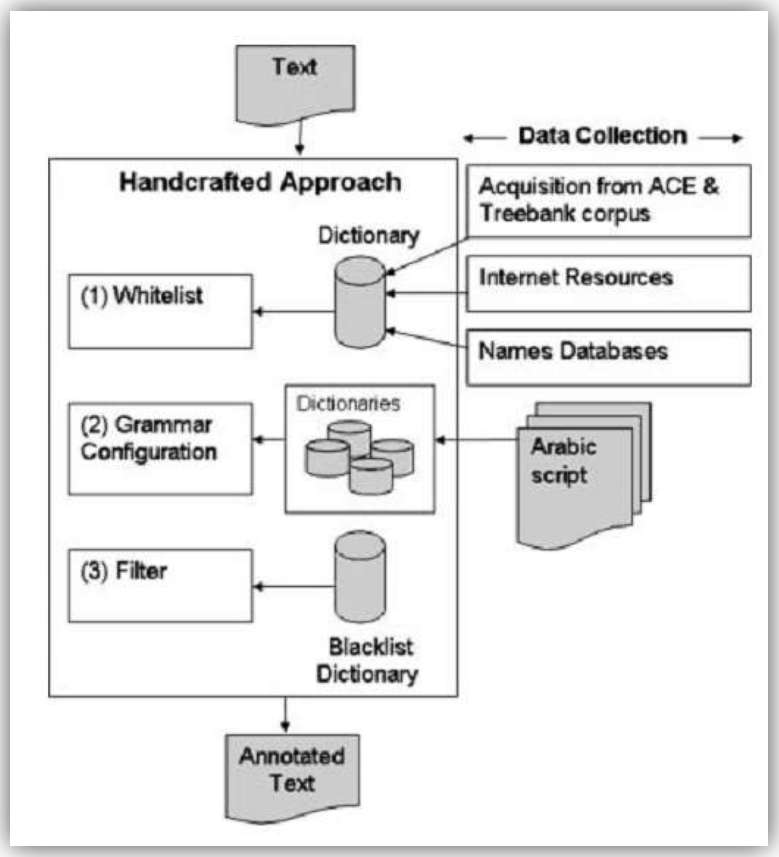


Fig. 3 Architecture of NERA system
(Source: [Shaalan & Raza, 2009])

So, in case of an intersection, "أحمد آباد" will be recognized as person name. In my opinion, there are some the drawbacks in "NERA" system when using this rule; where the phrase "أحمد أباد" can be a location but in this case the system can't retrieve the correct answer!. So, we suggest using the probabilities and the context surround the NEs for improving this linguistic rule and the similar rules.

Also, there are some the ways that can improve the performance and enhance the accuracy such as using large corpora and large dictionaries, and using Arabic text with error-free spelling.

5.2. Machine Learning (ML)-Based NER

The idea in ML-based approach is to study the features of the Named Entity classes in a large training corpus. There are two main components are used in the ML-Based systems, as following: (1) Large annotated corpora and (2) a probabilistic representation of the training data (A statistical model). Decision Trees, Support Vector Machines (SVM), Maximum Entropy (ME), Conditional Random Fields (CRF) and Hidden Markov Models (HMM) are examples of the statistical models.

5.2.1. The advantage and disadvantage of this approach

The advantage of this approach is that the use of ML approaches reduces the human effort needed for building a set of rules and gazeteers. In some cases, the ML-based approach is more flexible than the rule based approach.

The disadvantage of the machine learning approach is the needing to large corpora of annotated text. And this problem appears highly in Arabic NER because of the lack of linguistic resources.

5.2.2. Example on using this approach

(Benajiba, et al., 2007) built a NER system called "NERsys" for Arabic texts based-on n- grams and maximum entropy. The authors depend on their own testing and training corpora and gazeteers (ANERcorp & ANERgazet)[4].

First, we would to know how to use maximum entropy

Maximum entropy (ME) is "a general technique for estimating probability distributions from data" [Benajiba, et al., 2007]. The Maximum entropy calculates the best probability distribution according to the defined information. The following example will explain "how the ME classifier performs":

> *"Sudan's **Darfur** region remains the most pressing humanitarian problem in the world"*

If we want to classify the word "***Darfur***" as one of these classes (person, location, organization, or other)

- If we assumed that we don't have any information on the word "***Darfur***" then the distribution of the probability on the four classes

[4] Available at: http://www1.ccls.columbia.edu/~ybenajiba/downloads.html

will be the same. P(person)=P(Location)=P(Organization)=P(other) = ¼= 0.25.

- But, if we assumed that we have information for "***Darfur***" that it a word that begins with "Capital letter" and isn't "in the start of the sentence".

  - By using this information we will guess that "***Darfur***" is a proper name (one of those: Person, Location, or Organization) and it isn't any other NE.
  - So, distribution of the probability will be different. The most probability will be assigned to the 'Person', 'Location', and 'Organization'. P(person)=P(Location)=P(Organization)= 1/3 = 0.3.

In (Benajiba, et al., 2007), the results showed that using the maximum entropy can enhance the performance of the Arabic NER task without using any POS-tag information or text segmentation.

Table 7 shows a Comparison between Rule-Based and ML-Based

Table 7 Comparison between Rule-Based and ML-Based

	Rule-Based	**ML-Based**
Language Specialists	It depends on linguistic rules "hand-constructed rules" that require language specialists	It doesn't need language specialists
The Required Training Data	Small amount of training data	Large amount of annotated training data (very large corpora)
Time Consuming	Very time consuming	Automated
Changes	Some changes may be hard	Some changes require re-annotation of the entire training corpus
Quality	High quality	Less quality

5.3. Hybrid NER

The idea in the hybrid NER systems is combining the rule-based approach with the ML-based approach to overcome on the problems in both methods. Using the hybrid Arabic NER can improve the Arabic NER task.

5.3.1. The advantage and disadvantage of this approach

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

5.3.2. Example on using this approach

(Abdallah, et al. 2012) proposed a simple method for integrating the Machine learning-based approach with rule-based approach for Arabic NER as it is shown in Fig. 4.

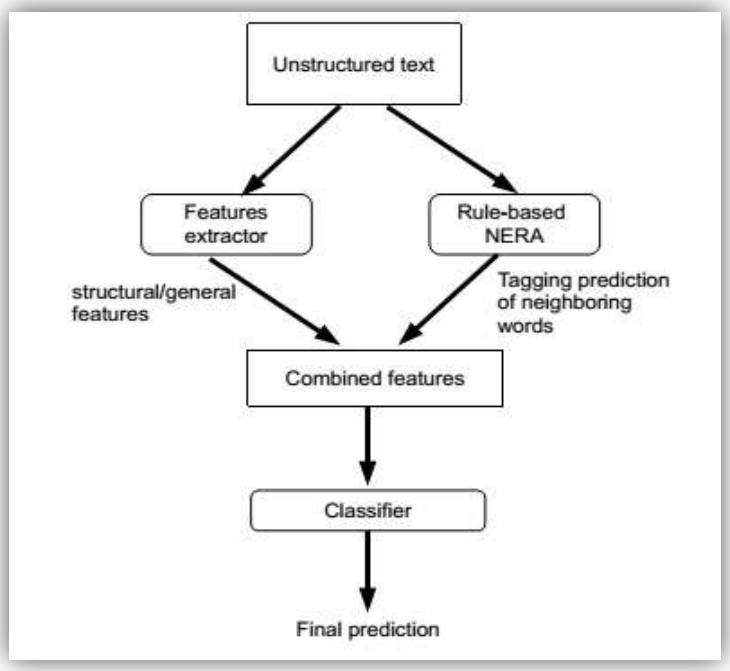


Fig. 4 Block diagram illustrating the integration between the Rule-Based and ML-Based (Source: [Abdallah, et al. 2012])

The authors focused on only three named entities (person name, location, and organization).

For training and testing the proposed method, they used two annotated corpora:

(1) The ACE 2003Multilingual Training Set[5].
(2) ANERcorp Corpus[6].

Steps to build their rule-based system:

1. Performing the recognition based on a dictionary lookup that containing lists of known named entities.
2. Using a parser, based on a set of grammar rules (represented as regular expressions).

Steps to build the proposed integrated approach:

1. Using the Stanford POS Tagger[7] to compute some of the general features such as word category and affixation that are defined as machine learning features.
2. Complementing the rule-based features with the other extracted features
3. Feeding all combining features to a decision tree classifier.

In (Abdallah, et al. 2012), the results of proved that the proposed hybrid approach is better than the pure rule-based system or the pure machine-learning classifier.

6. Evaluating the performance of the NER systems

The aim of the evaluation is to ensure if the NER system can enhance the performance or no and with which degree according to the used datasets. NER systems are evaluated by "running them on human-labeled data and comparing their results against this gold-standard" [Mohit, 2014]. If there are standard evaluation corpora, it will be easy to compare between the existed NER systems. Some researchers used the annotated datasets where every NE has a type assigned to it such as ("ANERcorp" and "ANERGazet"), and the evaluation measurements (precision, recall,

[5] Available to BUID under License from https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications
[6] Available for download from http://users.dsic.upv.es/ybenajiba/
[7] available at http://nlp.stanford.edu/software/stanford-postagger-2010-05-26.tgz

f-measure and accuracy) for evaluating their NER systems [Benajiba et al., 2007].

In this section, we presented the evaluation measurements and the evaluation metrics that are used for evaluating the performance of the Arabic NER systems.

6.1. Evaluation Measurements

The evaluation measurements such as "Precision, Recall, F-measure, and Accuracy" are the most used of the evaluation techniques [Al-Jumaily et al., 2012]. The equations of each measure will be defined in terms as in the following table:

Table 8 Terms in the evaluation measures equations

Term	Stands for	Definition
TP	((True-Positives))	"It counts the tokens correctly assigned to this category"
FP	((False-Positives))	"It counts the tokens incorrectly tagged to this category"
FN	((False-Negatives))	"It counts the tokens incorrectly rejected from this category"
TN	((True-Negatives))	"It counts the tokens correctly rejected from this category"

Precision

The precision is "the ratio of the retrieved tokens which are relevant in the corpus, i.e., it evaluates the exactness of the system" [Jumaily et al., 2012].

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Recall

Recall is "the ratio of the retrieved relevant tokens. It measures the ability of the system to retrieve a complete set of the relevant tokens from a corpus" [Jumaily et al., 2012].

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

F-measure

F-measure evaluates the effectiveness of the system.

$$Fmeasure = \frac{2 * (\Pr cision * Recall)}{(\Pr cision + Recall)} \quad (3)$$

6.2. Evaluation Metrics

There are three main NER scoring metrics used as part of NLP conferences: (1) **M**essage **U**nderstanding **C**onference "MUC", (2) **Co**mputational **N**atural **L**anguage **L**earning Conference "CoNLL", and (3) **A**utomatic **C**ontent **E**xtraction "ACE".

6.2.1. MUC Evaluations[8]

"MUC" is an evaluation metrics where a system is scored on two axes:

1. Its ability to find the correct type of the NE (TYPE).
2. Its ability to find the boundaries text surrounds the NE (TEXT).

The advantage of this method is taking into account all possible types of errors. But the ambiguity in the boundaries can lead to a problem.

6.2.2. CoNLL Evaluations[9]

"CoNLL" provides an exact match evaluation, where the entity is considered correct if it exactly matches the same type and text. This method is simple in calculating and analyzing results [Shaalan, 2014]. In this method, the Precision, recall and F-measures are used to calculate performance.

6.2.3. ACE Evaluations[10]

"ACE" evaluation deals with several kinds of errors into an integrated scoring mechanism where each type of error and each type of entity has different weight. Compared by "MUC" and "CoNLL" methods, "ACE" is more complex. Because of the complexity of this evaluation method, most studies used the "CoNLL" method for the evaluation [Alotabi, 2015].

According to the literature, we concluded that the **CoNLL** method can be a standard method for evaluating the Arabic NER.

[8] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/muc_sw/muc_sw_manual.html
[9] http://universaldependencies.org/conll17/evaluation.html
[10] https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications

7. Conclusion and Future Work

The Named Entity Recognition task has an importance in each field of NLP fields. In this research, we studied the importance of the NER task in some of these fields. Also the challenges that faced the NER task when using it with the Arabic language and the solutions that are used in the literature for each challenge were studied and analyzed. We founded that "analyzing the context surrounding the NEs" is the solution for the most of these challenges.

The studies on the Arabic NER used some the linguistic resources such as Corpora and Gazetteers and they also used some tools that support the Arabic language such as BAMA and MADAMIRA. Some of these resources and tools are freely available and the others are available under paid license agreements.

There are three approaches of the NER: (1) Rule-Based approach, (2) Machine Learning approach and (3) Hybrid approach. According to the literature the Hybrid-based can be the best approach in some cases.

NER systems are compared based on standard evaluations such as "MUC", "CoNLL", and "ACE"; and standard evaluation measurements such as "Precision", "Recall", and "F-measure". According to the literature, we concluded that the "CoNLL" Matrix was the most used because of its simplicity in the calculating and the analyzing the results. To the best of our knowledge, there isn't research that evaluates and compares the performance of the available Arabic NER systems.

In the future, we would to make a comparative study between using many statistical models (SVM, HMM, Maximum Entropy, CRF, etc.); and to detect the role for each for improving the Arabic NER task.

References

Abdallah, S., Shaalan, K., & Shoaib, M. (2012). Integrating rule-based system with classification for Arabic named entity recognition. In Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing),Springer- Verlag, Berlin Heidelberg, pages 311-322.

Al-Jumaily, H., Martínez, P., Martínez-Fernández, J. L., & Van der Goot, E. (2012). A real time Named Entity Recognition system for Arabic text mining. Language Resources and Evaluation, 46(4), 543-563.

Alotaibi, F. (2015). Fine-grained Arabic named entity recognition (Doctoral dissertation, University of Birmingham).

Benajiba, Y., Rosso, P., & Benedíruiz, J. M. (2007, February). Anersys: An arabic named entity recognition system based on maximum entropy. In International

Conference on Intelligent Text Processing and Computational Linguistics (pp. 143-153). Springer Berlin Heidelberg.

Dandashi, A., Al Jaam, J., & Foufou, S. (2016). Arabic Named Entity Recognition—A Survey and Analysis. In Intelligent Interactive Multimedia Systems and Services 2016 (pp. 83-96). Springer International Publishing.

Kaddoura, H. (2010). ALNER: Arabic location named entities (Doctoral dissertation, The British University in Dubai (BUiD)).

Mohit, B. (2014). Named entity recognition. In Natural Language Processing of Semitic Languages (pp. 221-245). Springer Berlin Heidelberg

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.

Saad, M. K., & Ashour, W. (2010). Arabic morphological tools for text mining. In: EEECS10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science. European University of Lefke, Cyprus, pp. 112–117.

Shaalan, K. & Raza, H. (2009). NERA: Named entity recognition for Arabic. Journal of the American Society for Information Science and Technology, 60(8), 1652-1663.

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

Image Encryption Scheme with Hashed Biometric Key

Ali M. Meligy[1], Hossam A. Diab[2] and Marwa S. ElDanaf[3]

ABSTRACT

Communities have always interested for the security of information over the eras. This has produce to the improvement of several information security methods, prominent among them is cryptography. In this paper we suggest a technique for image encryption that has fast execution speed and high level safety. The design tools of our algorithm are depend on a chaos-based feedback cryptographic scheme using the tent map and an external biometric secret key of 160-bit using hash function (SHA-1). Accordingly, the created external key is utilized to extract the initial seeds of the used chaotic mapping. We will see soon, the experiences show the efficiency of the suggested scheme in addition to its vulnerability to secret key changes and its resistance to various kinds of threats.

Keywords

Stream Cipher, Chaos Cryptography, Hash Function, SHA-1 Hash Algorithm, Tent Map, Logistic Map, Arnold Cat Map, Statistical Tests, Statistical analysis, Security analysis.

1. INTRODUCTION

Historically, the security of data was accomplished by a collection of physical security methods and trust. Stamps and subscriptions were confirmed of substantiation of legitimacy, this is appeared in the royal seals and signets by last emperors. Confidentially was accomplished by closing up documents and lockers. Integrity of information was fundamentally based on trust. Devoted manpower such the royal postal service was entrusted with the task of transferring information. The degree of information security was based on the trouble faced in compromising these security methods and breaking confidence. In the information age, though the demand for information security yet stays the same. Many of these physical security techniques are not only inadequate, but many are impracticable. The networked age has raised the danger of information safety probability of accumulation assets. Vulnerable data establishing on unlock and un-trusted networks (e.g. the Internet) can be simply accessed, duplicate or altered [3]. In impose to counter the security threat in the new age; a distinguished technique that has improved is encryption. The subject of encryption is called cryptography.

A hash function is a utility that takes a comparatively arbitrary quantity of input and makes an output of fixed size [6]. This concept feature makes them important in data structure, checksum algorithms for error detection, digital signature in information security etc. Cryptographic hash functions have another merit that is beyond hash functions - it is very rough to discover two distinct inputs that make the same output. This property gives a high level of certitude though not ultimate that several input values would product different output signature in most of the cases. This make cryptographic hash functions the hash functions that are used in information security related applications. SHA-1 (Secure Hash Algorithm) is one of hashing algorithms; it is used to generate the hashing value. It creates the hash value of 160 bits (20 bytes). It has the 80 number of rounds. The user which has the hash value can adjust the data. The hashing algorithm supply authenticity and integrity. If any user alters the data then the hash value will be varied [1].

[1]Prof.Dr of Computer science Faculty of science–menoufiya University–Mathematics department meligyali@hotmail.com
[2]Assistant Professor of Computer Science Faculty of science– Menoufiya University– Mathematics department hossamdiab_86@yahoo.com
[3]Researcher in Academy of Scientific Research and Technology marwa277.saeed@gmail.com

According to [7], Cryptographic Hash functions are one of the most significant tools in the field of cryptography and are used to realize a number of security goals. In [8] the essential characteristics of the Hash function that allow them to withstand to a satisfactory level all known cryptanalytic attack was highlighted.

In [2] mentioned to collision resistance property as collision freeness or strong collision resistance, second pre-image resistance as weak collision resistance and preimage resistance as one-wayness.

In [4] categorized collision resistance as the strongest property of all three, hardest to satisfy and easiest to breach, and breaking it is the goal of most attacks on hash functions.

In [5] the notion of the hash function security was expanded. In this extension they realized seven several security notions, three construct on pre-image resistance, three based on second pre-image resistance and one on collision resistance. Their work is based on public concept of hash function family that is a limited set of hash functions with common domain and range.

Similarly some chaos – based cryptosystems are used to solve the privacy and security problems of biometric templates. The secret keys are randomly generated and each session has different secret keys. Thus biometric templates are encrypted by means of chaotic cryptographic scheme which makes them more difficult to decipher under attacks [11]. Most properties are related to some requirements such as mixing and diffusion in the sense of cryptography. Therefore, chaotic cryptosystems have more useful and practical applications.

One of the simplest chaos functions that have been studied recently for cryptography applications is the logistic map. The logistic map function is expressed as:

$$x_{n+1} = r\, x_n(1 - x_n) \qquad (1)$$

Where x_n takes values in the interval [0, 1] and r € [0, 4]. It is one of the simplest models that present chaotic behavior [12].

Also, the proposed cipher utilizes the chaotic Tent map which can be depicted as follows:

$$f_\mu(x_{n+1}) = \begin{cases} \mu\, x_n & for\ x_n < 0.5 \\ \mu\,(1 - x_n) & for\ 0.5 \le x_n \end{cases} \qquad (2)$$

For the chaotic Tent map [12], the control parameter μ =2 yields a chaotic sequence x_n ϵ [0, 1] with a random behavior.

Also, popularized Arnold map in 1960, Russian mathematician Vladimir Arnold used the most general two-dimensional chaotic map for an image [9]; the name was Arnold Cat Map. If a matrix N * N, pixel with coordinates (x, y), we wrote to Arnold will be:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & p \\ q & pq+1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} (\mathrm{mod} N) \qquad p, q=1 \qquad (3)$$

Where q and p are real numbers, the (x', y') is the new position of original pixel (x, y). Since there only exists a linear transformation and mod function, it is very efficient to mix the pixel positions using the Arnold cat map. After several iterations, the correlation among the adjacent pixels can be muddled totally.

In this paper, present a technique for image encryption that has fast execution speed and high level safety. The design tools of our scheme are depend on a chaos-based feedback cryptographic scheme using the tent map and an external biometric secret key of 160-bit using hash function SHA-1. Accordingly, the generated external key is employed to derive the initial seeds of the applied chaotic mapping. Moreover, the

pixels are masked based on an iterative module which exploits a data-dependent feedback mechanism to mix the current cipher conditions with the previously masked pixels to get the encryption results.

The paper is organized as follows: Section 2 presents the structure of the proposed image encryption scheme. Experimental tests and numerical computations to emphasize the encryption quality of the presented scheme are suggested in Section 3. Section 4 debates the diverse security analyses of the presented scheme including statistical analysis and sensitivity analysis related to key and plaintext changes. Finally, Section 5 drafts the conclusions.

2. PROPOSED IMAGE ENCRYPTION SCHEME

The suggested cipher is a symmetric key stream cryptography algorithm in which three essential functions (the key expansion, encryption and decryption modules) are utilized by the dispatcher and recipient to acquire the encrypted and decrypted image, severally. First, a secret biometric image is utilized by the dispatcher/recipient to create the secret key by applying hash function (SHA-1). The encryption operations are utilized to the plain image to obtain the cipher image. The structure of the suggested cipher depends on a feedback mechanism in which the encryption of each pixel is made dependent on the encryption merits of the preceding cipher pixel, which in turn, makes the cryptosystem powerful against any kind of threats.

The next subsections demonstrate the three phases of the suggested scheme.

2.1 Key Expansion by Using SHA-1

In view of the fundamental wants of cryptology, the cipher text should be strongly related to the secret key and the security of the encryption algorithm only relies on obscuring this key. Moreover, the cipher should be strictly sensitive to little changes of the secret key. Thus, the strategy of randomly generating the key ensures these requirements. The proposed mechanism for key scheduling utilizes from a selected biometric image to obtain the desired key by using SHA-1.

The steps for key generation can be represented as follows:

Step 1: Input a biometric image Q.

Step 2: Transform biometric image Q to be square matrix m × m then Exchanged biometric image Q by using Arnold cat map Eq. 3, the existing matrix is titled by A.

Step 3: Produce matrix B with same length of A by utilizing Logistic map which mentioned in Eq. 1, where $X_0 = (\frac{\text{mean(A)}}{255}) \bmod 1$.

Step 4: However, the periodicity of Arnold cat map should degrade the security, because the possible threats may iterate the Arnold cat map incessantly to reemerge the original image. As a remedy, we modify the pixel values next to increase efficient key by BitXoring A and B.

$$C = A \oplus B \qquad (4)$$

Step 5: Derive 160 bit secret key by utilizing SHA-1 which take C in step 3 as input.

$$\text{SecretKey} = \text{SHA-1}(C) \qquad (5)$$

The test suite from NIST [10] was chosen to check the randomness of the sequence (secret key) produced by the suggested key expansion mechanism. This suite consists of a set of tests. Each test is independently applied to an n bits sequence (the same sequence in each test) to get a P-value. Particularly, the statistical

package consists of 16 tests [10]. These tests are performed on our proposed key generator and the obtained results are brief in Table 1. The estimations values confirm that the proposed generator can pass many of the underlying statistical tests and the basic requirements for the uniform distribution are met. Thus, the generated key stream is uniformly distributed and cannot be predicted by an enemy.

Table 1: Results of Statistical Tests NIST

No.	Statistical tests	P_Value	No.	Statistical tests	P_Value
1	Frequency (Monobit)	0.7518	9	Maurer's Universal, (L=7,Q=1280)	0
2	Block Frequency (M =20)	0.4224	10	Lempel-Ziv Compression	0
3	Runs Test	0.1522	11	Serial	0
4	Longest Runs of Ones(M=5)	0.0167	12	Approximate Entropy	0
5	Binary Matrix Rank(M = 7)	1.1872	13	Linear Complexity (M = 22)	0.1736
6	Spectral DFT	1.000	14	Cumulative Sums (Backward) Zero & One	0.7200& 1.000
7	Non-overlapping, M=12,B= [1 0 1]	0.0153	15	Random Excursions	0.6126
8	Overlapping, M =22,B =[1 1 0 1 1]	0.7856	16	Random Excursions Variant	0.4319

2.2 Proposed Encryption Scheme

The image encryption process utilizes an external biometric secret key of 160-bit long from SHA-1. Further, the secret key is divided into blocks of 8-bit. Now, the proposed scheme generates the initial seed of the employed chaotic tent map, $y_0$and the initial cipher pixel $C_0$from the extracted external biometric key K. Presume that the external secret key is represented as follows:

$$K = K_1\, K_2\, \ldots\, K_{20} \qquad (6)$$

Where K_i performs a block of 8-bit of the overall 160-bit biometric key K.

To compute y_0 and C_0, the following two steps are carried out:

$$y_0 = \left(\sum_{i=1}^{length(K)} K_i + \frac{K_{20}}{256}\right) mod\ 1 \qquad (7)$$

$$C_0 = \left(\left(\bigoplus_{i=1}^{length(K)} K_i + K_1/256\right) * 10\text{^}4\right) mod\ 256 \qquad (8)$$

The suggested encryption handles the plain image as a stream of pixels, each pixel is represented by 8-bit, and encrypts the input image pixel by pixel according to the following steps:

Step 1: Convert the 2D plain image P into 1D vector by reading the pixels from top left to bottom right sides. The obtained plain image vector and its corresponding cipher vector are denoted by P and C, respectively.

$$P = P_1 P_2\, \ldots\, P_m \qquad (9)$$

$$C = C_1 C_2\, \ldots\, C_m \qquad (10)$$

Step 2: Encrypt the current pixel P_i to obtain its corresponding cipher pixel C_i according to

$$C_i = [P_i \oplus \{([K_i + T(y_i)] \, mod 256) \oplus C_{i-1}\}] \ggg T_i \quad (11)$$

Where y_i denotes the current input for the Tent map T depicted in Eq. 2 and can be calculated as follows:

$$T(y_i) = \sum_{i=1}^{length\ of\ tent\ map} y_{i-1} \quad (12)$$

If the next value obtained is within the subinterval between (0.2, 0.8), the iteration goes on until a desired number not be subinterval between (0.2, 0.8) then exit from tent map. After encryption of each pixel, we modify the K_i and initial value of the tent map y_0 as follows:

$$K_i = K(T(y_i) mod\ 19\ ) + 1) \quad (13)$$

$$y_{i+1} = ((((K(i-1) mod 20) + 1) + K((T(y_i) mod 19) + 1)/255)) mod\ 1 \quad (14)$$

Step 3: Set i=i+1 and apply the step 2 until all pixels are encrypted.

Step 4: Convert C to 2D array to obtain the final encrypted image.

2.3 Proposed Decryption Scheme

Decryption is very simple; the same pad is generated but this time un-merged with the ciphertext to retrieve the plaintext. The decryption module receives an encrypted image (cipherimage) and the same 160-bit biometric secret key is generated and returns the original image (plainimage). The decryption scheme applies the same steps with the replacement of the encryption mapping with the inverse mapping of Eq. 11.

3. EXPERMENTAL RESULTS

To demonstrate the efficiency of the proposed cipher, several experiments are performed on a set of biometric images downloaded from CASIA (Chinese Academy of science and institute of Automation) database [15]. Also, for the numerical evaluation of the encryption quality, the correlation coefficient (C.C) between the plainimage and cipherimage is estimated. Mathematically, C.C can be expressed according to [13, 14] as follows:

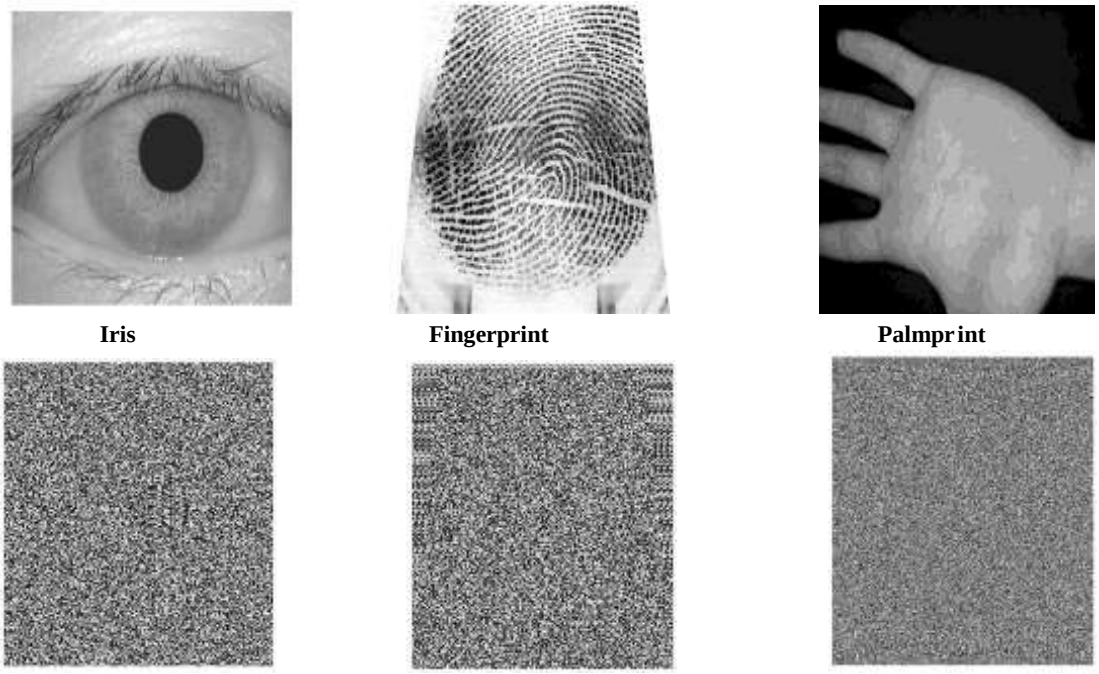
$$C.C = \frac{N \sum_{j=1}^{N}(x_j \times y_j) - \sum_{j=1}^{N} x_j \times \sum_{j=1}^{N} y_j}{\sqrt{(N \sum x_j^2 - (\sum_{j=1}^{N} x_j)^2) \times (N \sum y_j^2 - (\sum_{j=1}^{N} y_j)^2)}} \quad (15)$$

Where x and y denote the grey values of the pixels for the plainimage and the corresponding encryption result.

The encrypted images which are illustrated in Fig. 1 emphasize the feasibility of the proposed scheme. Obviously, the proposed scheme effectively conceals all features of the plainimage which means that the encrypted image is visually indistinguishable. Also, the results are compared with the standard encryption algorithms (AES, RC5, and RC6) in Table 2, where the proposed scheme retains the smallest Correlation Coefficients (C.C).

Table 2: The evaluation of encryption quality

Images	Correlation Coefficient			
	RC5	RC6	AES	Proposed Cipher
Iris001_1_1	0.0118	0.0225	0.0152	- 0.0073
Iris001_2_1	- 0.0258	0.0091	- 0.0884	0.0064
Fingerprint100_L0_0	0.0400	0.0078	-0.0119	0.0052
Fingerprint100_R0_0	0.0386	0.0056	0.0183	0.0031
Palmprint0001_m_L_01	- 0.0284	0.0067	0.0040	- 0.0044
Palmprint0001_m_R_01	- 0.0061	0.0050	0.0031	0.0032



Iris **Fingerprint** **Palmprint**

Iris_enc. **Fingerprint_enc.** **Palmprint_enc.**

Fig 1: Encryption by the propose scheme

4. SECURITY ANALYSIS

An acceptable encryption algorithm must thwart all kinds of cryptanalytic threats such as statistical attacks and exhaustive search attacks, differential attacks and related key attacks [13, 14, 16]. In this section, several security tests are applied to the proposed cipher to demonstrate its satisfactory security level.

4.1 Statistical Analysis

From cryptanalysis point of view, statistical analysis may enable an attacker to crack the cipher and recover the plain image from its cipherimage. Indeed, several cryptography schemes have been successfully broken through the statistical analysis such as permutation based ciphers. Hence, to confirm the strength of the proposed cryptosystem, the statistical analysis based on histogram and adjacent pixel correlations analysis are performed. The obtained results demonstrate the ideality of the proposed cipher with respect to statistical attacks.

4.1.1 Correlation of Two Adjacent Pixels

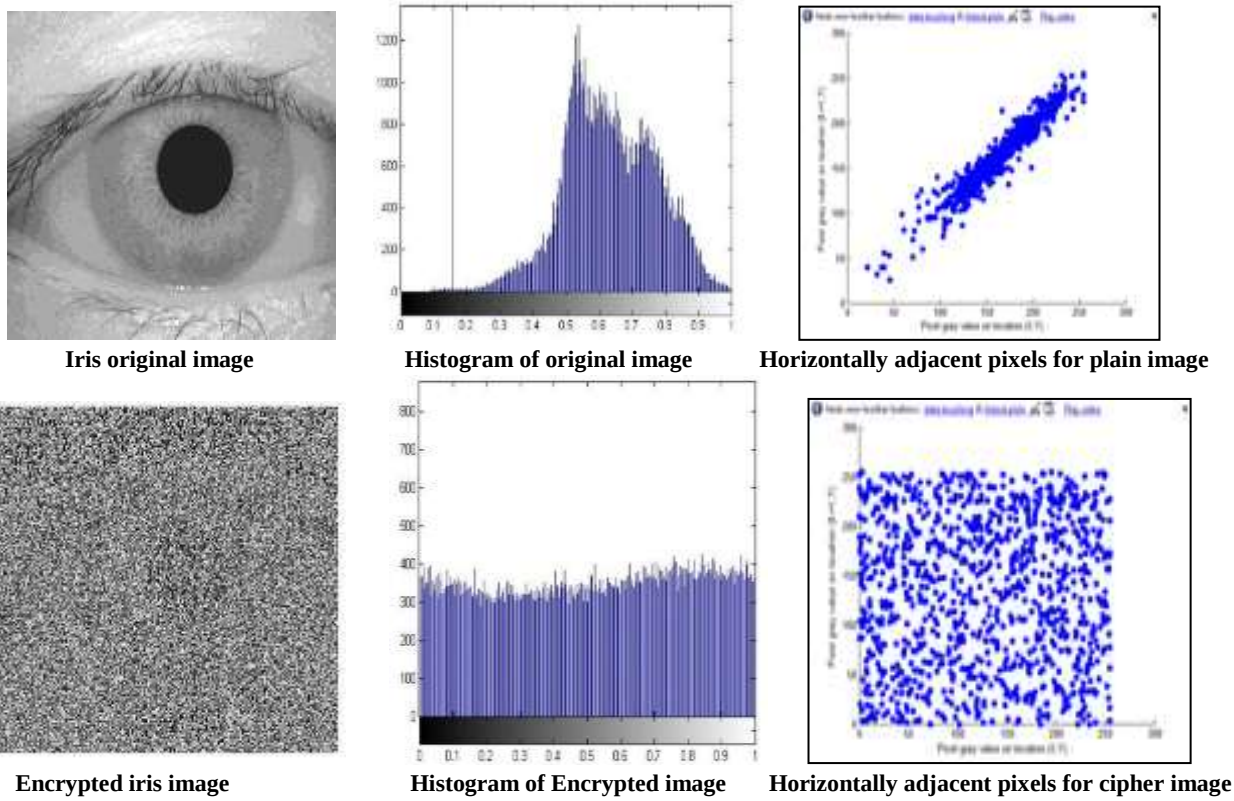
The correlations between neighboring pixels are tested for horizontal, diagonal, and vertical adjacent pixels for the plainimage and the associated cipherimage. First, several pairs of adjacent pixels in different directions are randomly selected. Then, calculate the correlation coefficient between them according to Eq. 15. The results of the adjacent correlation analysis for horizontal pixels for iris image and its related cipherimage are illustrated in Fig. 2. The obtained values for the correlation coefficients in the plainimage and cipherimage are tabulated in Table 3 for different directions. Obviously, there is an extraneous correlation between adjacent pixels in the cipherimage. On the other hand, the plainimage appears well correlated adjacent pixels which prove that the success of the proposed scheme in decreasing such correlation.

Table 3: Obtained Values of C.C between Adjacent Pixels for Plainimage/Cipherimage

Direction	Plainimage	Cipherimage
Horizontal	0.9701	0.0679
Vertical	0.9752	0.0009
Diagonal	0.9539	-0.0459

4.1.2 Gray Histograms Analysis

Additionally, to prevent an opponent from exploiting the statistical features of the cipherimages to obtain valuable information about the plain image, the cipherimage must bear a high dissimilarity to the original image. The histogram of several encrypted biometric images and its related original biometric images are studied. One of these examples shown in Fig. 2 displays the histogram of a cipherimage and the histogram of corresponding original image denoted by iris image. It is clear that the encipher image histogram is uniform distributed and notably dissimilar to the relevant histogram of the corresponding original image and consequently does not afford any indication about the original plainimage. Thus, an opponent cannot apply any statistical analysis on the proposed cipher.



Iris original image | **Histogram of original image** | **Horizontally adjacent pixels for plain image**

Encrypted iris image | **Histogram of Encrypted image** | **Horizontally adjacent pixels for cipher image**

Fig. 2. Histogram and correlation coefficient of two horizontally adjacent pixels in plain image and its related cipher image

4.2 Information Entropy Analysis

Information entropy is considered as a significant indicator to randomness degree. According to Shannon's theory [17], the entropy of an information source IS can be defined as follows:

$$H(IS) = -\sum_{i=0}^{2^L-1} P(IS_i) log_2 P(IS_i) \qquad (16)$$

Where P(**IS$_i$**) denotes the probability of symbol **IS$_i$**, and L is the number of bits used in representation of symbols of the source IS. According to this definition, it is found that the idea value of entropy for a random image with 28 (256) gray levels equals 8. To test the safety of the suggested scheme against the entropy attack, the entropy values for several images encrypted by the proposed scheme are estimated and are displayed in Table 4. The obtained estimators are too close to the expected value of 8 of a perfect random image. Thus, the proposed image cipher can defy the entropy attacks.

Table 4: Results of Information Entropy Analysis.

Images	Entropy	Images	Entropy
Iris 001_1_1	7.9977	Fingerprint 100_R0_0	7.9982

Iris 001_2_1	7.9982	Palmprint 0001_m_L_01	7.9995
Fingerprint 100_L0_0	7.9984	Palmprint 0001_m_R_01	7.9994

4.3 SENSITIVITY ANALYSIS

A perfect encryption procedure must be wholly sensitive to small modification in the associated encryption key and the original plainimage. Namely, the trivial variation on a single bit in either the plainimage or the secret key should yield a significant change in the cipherimage (i.e. completely different enciphered image). To verify the robustness of the proposed algorithm, the following analysis is employed.

4.3.1 Key Sensitivity Analysis

One aspect of key sensitivity for a secure cipher is the failure of restoring the plainimage from cipher image if there is only a slight diversity between decryption and encryption keys. Really, this feature also promises the high resistance of the cryptosystem to brute-force attacks. On the other hand, the resulting cipher image from a small changing in the encryption key must result in extremely different enciphered image. To test the high sensitivity of the proposed cipher to the changes in the secret key, the following steps are carried out:

1) Use the secret key K_1 to encrypt the plainimage shown in Fig. 4(a) and the ciphered image is denoted as A as depicted in Fig. 4(b).
2) The same plainimage is enciphered again by the key K_2 where this key is different from the previous key K_1 in only the most significant bit. The resulting image B is displayed in Fig. 4 (c).
3) Finally, the same image is enciphered again by the key K_3 where this key is different from K_1 in only the least significant bit. The resulting image C is illustrated in Fig. 4 (d).
4) Compare the enciphered images A, B and C to find their differences.

Fig. 4 shows the plainimage and the three corresponding cipher images produced from the applications of the aforementioned steps. To computationally compare theses encrypted images, the correlation between each pair of them is evaluated. Table 5 lists the obtained results for correlation. It is obvious that modifying only one bit of the secret key yields entirely distinct enciphered images with insignificant correlation between them.

Table 5: Obtained C.C Analysis for Cipher by Slightly Different Secret Key.

Image1	Image2	Correlation coefficient
Encrypted image A	Encrypted image B	- 0.0005
Encrypted image B	Encrypted image C	- 0.0017
Encrypted image C	Encrypted image A	- 0.0056

a) Original image **b) Encrypted image CI_1** **c) Encrypted image CI2** **d) Encrypted image CI3**

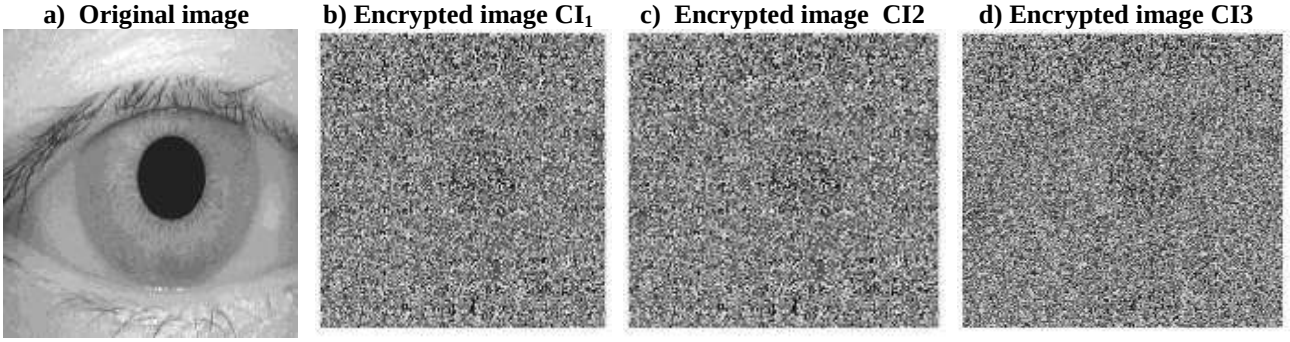
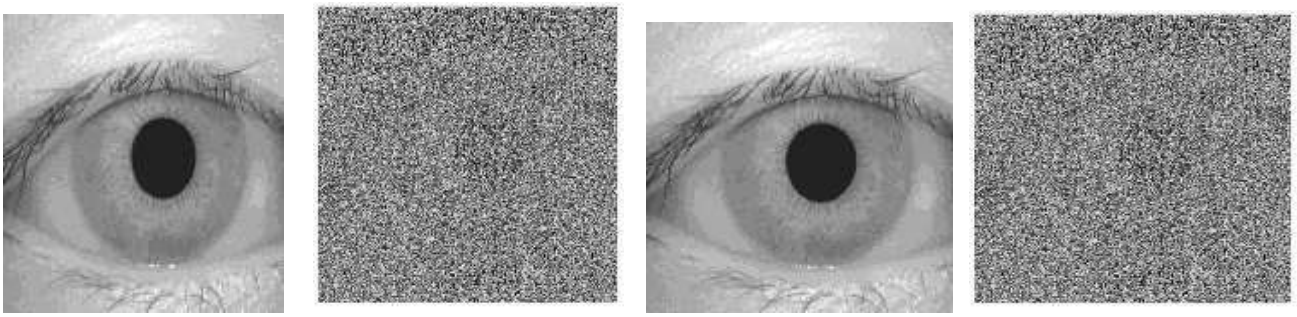


Fig. 4: Key sensitive

Moreover, the attempt to recover the original image from the enciphered one with slightly different key fails.

Particularly, Fig. 5 (a) and Fig. 5 (b) explain the original plainimage and the associated encrypted image produced using the secret key K1, Severally, whereas Fig. 5 (c) and Fig. 5 (d) draw the retrieved images from the decryption procedure with a correct key K1 and a slightly different key K2, respectively. Distinctly, the deciphering with a somewhat distinct key cannot succeed.



a) **Original image** | b) **enciphered image With secret key** | c) **Recovered image With secret key** | d) **Recovered image by slightly different key**

Fig. 5: Key sensitive test

4.3.2 Plainimage Sensitivity Analysis

Another demand for a perfect encryption technique is its high allergy to little changes of the original image. To evaluate the suggested algorithm in this trend, two standards can be applied. The first one is the Number of Pixels Change Rate (NPCR) and the second is the Unified Average Changing Intensity (UACI). Presume that P1 and P2 are two images with only one pixel various, and the secret key applied is indicated by K, the next proceedings are utilized:

1) Use the secret key *K* to encrypt the first plainimage P_1 displayed in Fig. 4(a) and the related enciphered image is denoted as C_1.
2) The same secret key *K* is used to encrypt the second plainimage P_2 which is different from P_1 in only one pixel. The resulting image is named C_2.
3) Finally, compute the values of UACI and NPCR according to the following equations[13,16]:

$$UACI = \frac{1}{W \times H}\left[\sum_{i,j}\frac{|C_1(i,j) - C_1(i,j)|}{255}\right] \times 100\% \quad (17)$$

$$NPCR = \sum_{i,j}\frac{D(i,j)}{W \times H} \times 100\,\% \quad (18)$$

Where *W* and *H* are the width and height of C_1 or C_2 and $D(i, j)$ is evaluated as follows:

$$D(i,j) = \begin{cases} 0 & if\ C_1(i,j) = C_2(i,j) \\ 1 & otherwise \end{cases} \quad (19)$$

The test is executed on iris biometric image of size 280 x 320, and the result of NPCR estimator is found to be over 99.76%. As well, the value of UACI estimation is calculated to be over 33.49%. The obtained values of NPCR and UACI demonstrate that the suggested cipher is robustly sensitive to modifications happened in the original image.

5. Conclusion

A prospective technique for encryption images is suggested which employs one of chaotic systems (tent map). The suggested is depended on a chaos- feedback mechanism. The major merit of the suggested scheme depends on the employment of biometric secret image extract an external secret key by utilizing

Arnold cat map for shuffling places of biometric pixels and then employs SHA-1 hash algorithm and logistic map to change the values of pixels, which enhance effective key. The encoding of each pixel depends on the secret key, the preceding enciphered pixel and the output of tent map. The performed empirical results explain that the encrypted image has tiny correlations between neighbor pixels, roughly uniform image histogram which can be seen as roughly random image. Furthermore, the plain image allergy parameters are near to their exemplary values of 99.76% and 33.49%. The suggested scheme has high plain-image sensibility. So the suggested scheme is resistive to statistical and differential threats. Thus, the suggested scheme can is powerful versus the joint assaults.

References

[1] J. Menezes, S. A. Vanstone, and P. C. V. Oorschot, " *Handbook of Applied Cryptography*", CRC Press, Inc., 1996.

[2] Gauravram, P. *"Cryptographic hash functions: Cryptanalysis, design and applications"*, Ph.D. thesis, Faculty of Information Technology, Queensland University of Technology, Brisbane, Australia. 2003.

[3] Kahn, D. "*The Codebreakers: The comprehensive history of secret communication from ancient times to the Internet"*, (Revised edition). Scribner. 1996.

[4] Lai, X., and Massey, J. L. *"Hash function based on block ciphers"*, In EUROCRYPT, 1992, pp.55-70, 1992.

[5] Rogaway, P., and Shrimpton, T. *"Cryptographic hash function basics: Definitions, implications and separations for preimage resistance, second preimage resistance, and collision resistance"*, in FSE, pp.371-388, 2014.

[6] Silva, J. E. *"An overview of cryptographic hash functions and their uses"*, GIAC Security Essentials Practical Version 1.4b Option. 2003.

[7] Sobti, R., and Geetha, G. *"Cryptographic hash functions: A review"*, IJCSI International Journal of Computer Science Issues, 9(2), 461-479, 2012.

[8] In Wikipedia (2014), *"Information Security"*, Retrieved from, [online]. Available: http://en.wikipedia.org/wiki/Information_security

[9] R Jun, P., J. Shangzhu, and L. Yongguo, *"Design and Analysis of an Image Encryption Scheme Based on Chaotic Maps"*, International Conference ICICTA, pp. 1115-1118, 2010.

[10] National Institute of Standards and Technology. *"A statistical test suite for random and pseudorandom number generators for cryptographic applications"*, Special publication 800-22. Revision 1. 2008.

[11] Z.Yun-peng, and Z. Zheng-jun. *"Digital Image Encryption Algorithm Based on Chaos and Improved DES"*. Proceedings of the IEEE International Conference on Systems. Man, and Cybernetics San Antonio. TX, USA. 2009.

[12] Muhammad Khurram Khan, and Jiashu Zhang, *"Implementing Templates Security in Remote Biometric Authentication System"*, IEEE Conf. Proceedings on CIS'06, China. pp. 1396-1400. vol.2, 2006.

[13] G. Chen, Y. Mao, and C.K. Chui, *"A symmetric image encryption based on 3D chaotic maps"*, Chaos Solitons Fractals. vol. 21. pp. 749-761, 2004.

[14] Pareek ,N. K., Patidar , V., Sud, .K .K. , *"Image encryption using chaotic logistic map"*, In Image and Vision Computing 24-926–934. Elsevier, 2006.

[15] CASIA Iris Database . [Online March, 2009] http://sinobiometric.com.

[16] I. A. Ismail, M. Amin, and H. Diab, "A digital image encryption algorithm based a composition of two chaotic logistic maps", International Journal of Network Security. vol. 11. no. 1. pp. 1-10, 2010.

[17] Jiri Giesl, Ladislav Behal and Karel Vlcek, *"Improving Chaos Image Encryption Speed"*. International Journal of Future Generation Communication and Networking Vol. 2. No. 3, 2009.

[18] N.K. Pareek, Vinod Patidar, and K.K. Sud, *"Cryptography using multiple onedimensional chaotic maps"*, Commun. Nonlinear Sci. Numer. Simul. 10 (7) 715–723, 2005.

[19] Nitumoni Hazarika and Monjul Saiki, *"A Novel Partial Image Encryption using Chaotic Logistic Map"*. International conference on Signal Processing and Integrated Networks (SPIN), 2014.

Cauchy Based fuzzy neural network with Mutual Subsethood Product Inference

Nelly S. Amer [1], Hesham A. Hefny[2]

Abstract

This paper presents a mutual subsethood product fuzzy neural model based on Cauchy fuzzy sets. The proposed model has the ability to perform classification using both of numeric and linguistic inputs simultaneously. Fuzzy rule-based knowledge is translated into network architecture. Connections in the network are represented by Cauchy fuzzy sets. The firing degrees of the fuzzy IF-Then rule are based on fuzzy mutual subsethood similarity measure. Which is computed neither approximated nor numerically. It is computed by an exact formula. We focus on the classification ability of the model and demonstrate its performance on two benchmark classification problems: the IRIS data classification, Hepatitis medical diagnosis.

Keywords: fuzzy neural network, Fuzzy mutual subsethood.

1. Introduction

Fuzzy systems have been successfully used in a variety of applications, such as pattern recognition, automatic control and fuzzy inference systems [1].
It is very difficult to find a global function or analytical structure for a nonlinear system. In contrast, fuzzy logic provides an inference morphology that enables approximate human reasoning capability to be applied in a fuzzy inference system. Therefore, a fuzzy inference system employing fuzzy logical rules can model the quantitative aspects of human knowledge and reasoning processes without employing precise quantitative analysis.

In recent past, artificial neural network has also played an important role in solving many engineering problems. Neural network has advantages such as learning, adaptation, fault tolerance, parallelism, and generalization [2]. Fuzzy systems utilizing the learning capability of neural networks can successfully construct the nonlinear input output mapping for many applications. The term neuro fuzzy system [3] refers to an adaptive fuzzy systems, where neural network learning techniques are employed to adjust the shape of fuzzy sets, i.e. linguistic values, of the fuzzy rule base that constitute the model of the considered nonlinear phenomenon. Such a hybrid model has been applied efficiently in several applications including classification, function

[1] Dept. of computer sciences Institute of statistical studies and researches ,Cairo University.
[2] Dept. of computer sciences Institute of statistical studies and researches ,Cairo University.

approximation [4]. In the subsethood based fuzzy neural models, the adaptive fuzzy model can handle simultaneous admission of fuzzy or numeric inputs along with the integration of a fuzzy mutual subsethood measure for activity propagation

Fuzzy neural model based on mutual subsethood measure has a lot of applications. Several approaches have been proposed in literature to compute the firing degrees of the fuzzy IF-Then rule based on mutual subsethood similarity measure. Most research works made in this area adopt triangle or trapezoidal fuzzy sets due to the simplicity for obtaining an exact mathematical formula for the overlapping area. However, when dealing with continuous fuzzy sets, most of research works are based on numerical approximation, or approximate similarity formulas. Some researchers also used the possibility measure as a measure of similarity. In [5,6,7,8 and 9] the authors tried to calculate exact formula for similarity measure of Gaussian fuzzy sets, But these exact formulas were for four different possible cases of intersection, a general closed analytical formula could not be found, and hence high computational effort was done. Finally, in [10] , the generalized analytical formula of similarity measure of Gaussian fuzzy sets has been successfully obtained which can be used for distinguishability quantification. in [11], an exact analytical formula of mutual subsethood similarity measure of Cauchy fuzzy sets has been obtained which the work in this paper is inspired by it.

The organization of the paper is as follows: section two provides the architectural of the model. Section three describes the numeric and linguistic inputs. Section four presents the learning algorithm. Section five provides the set theoretic similarity measure for Cauchy fuzzy sets, and Section six demonstrates the application of the model pattern classification.

2. Architecture of the model

The proposed fuzzy neural model represents fuzzy rules of the form

If A_1 is SHORT and A_2 is TALL then B is MEDIUM.

Where A_1, A_2 , B , SHORT, MEDIUM, and TALL fuzzy are fuzzy sets defined, respectively, on input or output universes of discourse (UODs). As seen in fig.1. The input nodes of the first layer in the fuzzy neural network represent domain variables or features, and output nodes represent target variables or classes. Each hidden node represents a rule, and input-hidden node connections represent fuzzy rule antecedents. Each hidden-output node connection represents a fuzzy-rule consequent. Fuzzy sets corresponding to linguistic labels of fuzzy if–then rules (such as SHORT, MEDIUM, and TALL), are defined on

input and output UODs and are represented by Cauchy membership functions specified by a center and spread. Fuzzy weights w_{ji} from rule nodes j to input nodes i are thus modeled by the center w_{ji}^{a} and spread w_{ji}^{b} of a Cauchy fuzzy set and denoted by $w_{ji} = \left(w_{ji}^{a}, w_{ji}^{b}\right)$. By the same way, consequent fuzzy weights from output nodes to rule nodes are denoted by $w_{kj} = \left(w_{kj}^{a}, w_{kj}^{b}\right)$.
Subsethood Product Fuzzy Neural Inference system (SuPFuNIS) can simultaneously admit numeric inputs as well as fuzzy inputs. Numeric inputs are first fuzzified hence all inputs to the network are fuzzy. Since antecedent weights are also fuzzy, then it requires a method to transmit a fuzzy signal along a fuzzy weight. In our SuPFuNIS model fuzzy mutual subsethood is employed to handle Signal transmission along the fuzzy weight.

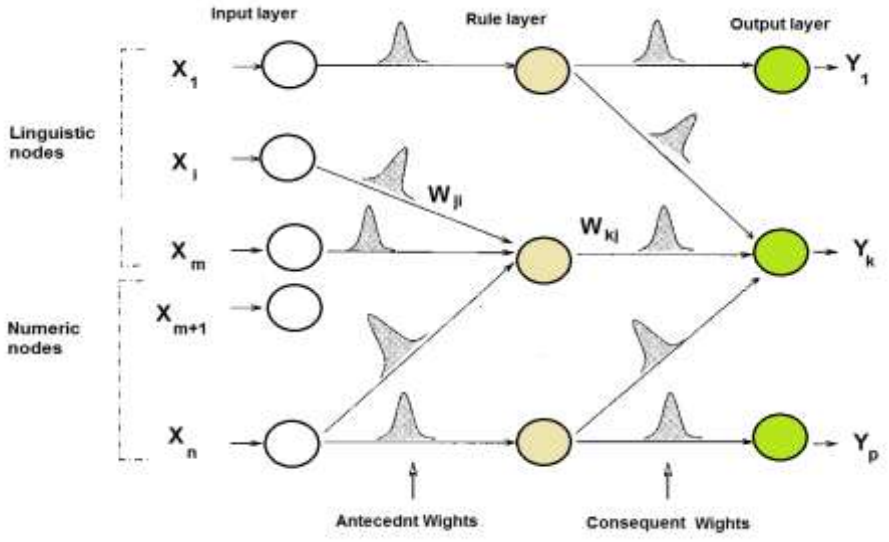


Fig. 1. Architecture of the fuzzy neural model.

3. Numeric and linguistic inputs

In this section we will show how the model can accept numeric inputs or ligustic inputs. It is known that The input features $(x_1, \ldots\ldots x_n)$ can be either numeric or linguistic. Linguistic nodes accept a linguistic input represented by a fuzzy set with a Cauchy membership function and modeled by a center a_i and spread b_i. The linguistic input feature x_i is represented by the pair(a_i, b_i). Numeric nodes accept numeric inputs and fuzzify them into Cauchy fuzzy sets. They are fuzzified by treating them as the center of a Cauchy membership function with a specified chosen spread. We chose a spread value of 0.5 for the applications presented in this paper. Therefore, the signal from a numeric node of the input layer also is represented by the pair(a_i, b_i). Antecedent connections uniformly receive signals of the form(a_i, b_i). Signals $S(x_i) = (a_i, b_i)$ are

transmitted to hidden rule nodes through fuzzy weights which is of the form(a_{ji}, b_{ji}).

4. The learning algorithm

The SuPFuNIS model is trained by back propagation supervised learning [9] . The model is trained by repeating presentation of a set of input patterns drawn from the training set. The firing degree of the fuzzy IF-Then rule will be computed set theoretic which will be shown in the next section. The output of the network is compared with the desired value to obtain the error, and network weights are changed on the basis of an error minimization criterion. Once the network is trained to the desired level of error, it is tested by presenting a new set of input patterns drawn from the testing set.

5. The Fuzzy mutual subsethood

The fuzzy mutual subsethood is the set theoretic similarity measure which is used to measure the similar degree between two fuzzy sets [12].

The set-theoretic similarity measure usually used in interpretability analysis is:

$$S(A,B) = \frac{|A \cap B|}{|A|+|B|-|A \cap B|} \quad ,where\ A\ and\ B\ are\ fuzzy\ sets. \qquad (1)$$

Where |.| denotes the cardinality of the set [13].
For the Cauchy fuzzy sets, first, the general form of Cauchy fuzzy set is

$$\mu(x) = \frac{1}{1+(\frac{x-a}{b})^2} \qquad (2)$$

where a is the center, b is the width and x belong to the domain

Let $\mu_A(x)$ and $\mu_B(x)$ be two Cauchy fuzzy sets of the following forms

$$\mu_A(x) = \frac{1}{1+(\frac{x-a_1}{b_1})^2} \quad , \qquad (3)$$

$$\mu_B(x) = \frac{1}{1+(\frac{x-a_2}{b_2})^2} \qquad (4)$$

The cardinality of Cauchy fuzzy sets calculation becomes integration as follows [11]:

$$|A| = \int_{-\infty}^{+\infty} \mu_A(x) = \int_{-\infty}^{+\infty} \frac{1}{1+(\frac{x-a_1}{b_1})^2} = b_1\pi \quad (5)$$

$$|B| = \int_{-\infty}^{+\infty} \mu_B(x) = \int_{-\infty}^{+\infty} \frac{1}{1+(\frac{x-a_2}{b_2})^2} = b_2\pi \quad (6)$$

$$|A \cap B| = b_{min}\pi + \Omega \quad (7)$$

Where

$$\Omega = (\mathrm{b}_{max} - \mathrm{b}_{min}) \tan^{-1}\left(\frac{a_{max} - a_{min}}{b_{max} - b_{min}}\right) - (b_{max} + b_{min}) \tan^{-1}\left(\frac{a_{max} - a_{min}}{b_{max} + b_{min}}\right)$$

$$b_{min} = \min(b_1, b_2)$$

$$b_{max} = \max(b_1, b_2)$$

$$a_{min} = \min(a_1, a_2)$$

$$a_{max} = \max(a_1, a_2) \quad (8)$$

The generalized formula for the similarity measure between two Cauchy fuzzy sets A and B with centers $a_1\ and\ a_2$ and widths $b_1\ and\ b_2$ can be obtained directly as follows:

$$S(A,B) = \frac{b_{min}\pi+\Omega}{b_{max}\pi+\Omega} \quad (9)$$

6. Applications in pattern

The pattern classification problem deals with categorization of an unseen pattern to 1-of- classes. These classes are either predefined or are learned based on the similarity of patterns. Common approaches for pattern recognition include neural networks, statistical methods, discriminant analysis, fuzzy systems, neural networks, and hybrid methods [5]. Attempts have also been made to design hybrid systems involving multiple models.

Our model in this paper has been tested on two different pattern classification benchmark data sets: IRIS data and hepatitis diagnosis [6]. For the IRIS classification problem, we report the resubstitution error which is uses the same data for training and testing and it is computed as the number of misclassification of patterns of all classes. For the hepatitis diagnosis

classification problems, we report the test errors where the network is trained on a training data set that is distinct from the test set. The test error is the number of misclassification of patterns of all classes.

6.1 IRIS Data Classification

IRIS data involves classification of three classes of the IRIS flower namely, IRIS sestosa, IRIS versicolor, and IRIS virginica on four feature of the IRIS flower:

1) Sepal length;
2) Sepal width;
3) Petal length;
4) Petal width.

There are 50 patterns (of four features) for each of the three subspecies of IRIS flower. The input pattern set thus comprises 150 four-dimensional patterns. This data is available from UCI repository of machine learning databases from http://www.ics.uci.edu/~mlearn/MLRepository.html.

The input layer consists of four numeric nodes and output layer comprises three class nodes. The number of nodes in the rule layer depends on the number of rules for which the network is trained. To train the network, initially, the centers of antecedent weight fuzzy sets were randomized in the range of the minimum and maximum values of respective input features of IRIS data. These ranges are (4.3000, 7.9000) which represent the minimum and the maximum of sepal length of 150 IRIS patterns, (2.0000, 4.4000) which represent the minimum and the maximum of sepal width of 150 IRIS patterns, (1.0000, 6.9000) which represent the minimum and the maximum of petal length of 150 IRIS patterns, and (0.1000, 2.5000) which represent the minimum and the maximum of petal width of 150 IRIS patterns. The centers of hidden-output weight fuzzy sets were randomized in the range (0, 1), and the spreads of all fuzzy weights were randomized in the range (0.2, 0.9). The feature spreads are taken as 0.5. All 150 patterns of the IRIS data were presented sequentially to the input layer of the network for training. The learning rate and momentum were both taken as 0.0007 and kept constant during the training period.

Once the network was trained, the test patterns (which again comprised all 150 patterns of IRIS data) were presented to the trained network and the resubstitution error computed.

TABLE I
BEST RESUBSTITUTION ACCURACY FOR IRIS DATA FOR DIFFERENT SOFT COPMTING ALGORITHMS

Method	Rules	Resubstitution accuracy (%)
FuGeNeSys	5	100
NEFCLASS	7	96.7
ReFuNN	9	95.3
EFuNN	17	95.3
FuNe-I	7	96.0
SUPFUNIS	5	100
Our Model	3	100

Table I, [8] compares between our model and other techniques according to resubstitution accuracy which indicates that our model can strongly classify IRIS data with least number of rules.

6.2 Medical Diagnosis

This problem deals with hepatitis diagnosis which requires classifying patients into two classes Die or Live on the basis of features which are both numeric and linguistic. The data can be obtained from http://www.ics.uci.edu/~mlearn/MLRepository.html.

The hepatitis data set has 155 patterns of 19 input features with a number of missing values. There are six numeric features namely Age, Bilirubin, Alk Phosphate, SGOT, Albumin, and Protime, and the remaining 13 features are linguistic in nature.

As there are a number of missing data, preprocessing of data is required. The data set consists of 75 patterns that have one or more features unspecified. A new set of data was formed by fitting some of the missing numeric values. Twenty patterns which had either a missing symbolic feature value, or more

than two missing numeric feature values were first discarded. The missing numeric values in the remaining 55 incomplete cases were filled with the average value of the missing feature calculated on a class-wise basis from the 80 original complete data [5 and 6]. By this method we were able to reconstruct a data set of 135 patterns. The numeric features of these 135 patterns were normalized feature-wise in the range [0, 1]. Symbolic features(yes/no or male/female) were represented by constructing two fuzzy sets: the symbolic value "no" represented by a fuzzy set with Cauchy membership function having center as zero and spread as 0.5, and "yes" represented by a Cauchy membership function centered at one and spread 0.5. The spreads were assumed to be trainable during the learning procedure.

Experiments were conducted using two data sets: Data Set 1 comprising of only 80 of 155 patterns that were originally complete in all respects; and Data Set 2 comprising 135 patterns (80 originally complete and 55 reconstructed). For training, 70% patterns were randomly chosen and the remaining 30% were used for testing. Five combinations of such 70 %(train) and 30 %(test) were randomly generated separately for Data Set 1 and Data Set 2. Experiments were then conducted on each of these individual data set combinations using SuPFuNIS architecture which had 19 input nodes, 3 nodes in the hidden layer and 2 nodes in the output layer. During the training, both learning rate and momentum were kept constant as 0.0007.

Table II shows the average classification accuracy obtained using SuPFuNIS for both data sets, compared with Kumar [8] to solve the same problem and it refers to the testing accuracy is raised in our model where the average of testing accuracy of data set 1 in our model is 94 % where in Kumar was 91.67. In data set 2 our testing accuracy average is 97 where in Kumar model was 96.5.

TABLE II
TESTING ACCURSCY IN % USING THREE RULES FOR HEPATITIS DATA

Method	Experiment	1	2	3	4	5	Average %
Kumar Model	Data Set1	91.67	100	87.5	87.5	91.67	91.67
	Data Set2	97.5	97.5	97.5	92.5	97.50	96.50
Our Model	Data Set1	95.83	95.83	95.83	91.67	91.67	94
	Data Set2	97.5	95	97.5	97.5	97.5	97

Conclusion

In this paper we proposed a SuPFuNIS model which can handle both numeric and linguistic inputs and we showed that it was different of other SuPFuNIS models where first it used Cauchy function membership and a lot of other model used Gaussian membership. Second it computed the mutual subsethood by exact formula not approximated, not numerically and not exact with different cases for intersections as Kumar. And hence the model was low computationally and its efficiency appeared in the two patterns classification cases.

References

[1] Song Hengjie et al, "a fuzzy neural network with fuzzy impact grade", Neurocomputing 72, pp 3098–3122, April 2009.

[2] Hengjie Song et al, "Implementation of Fuzzy Cognitive Maps Based on Fuzzy neural Network and Application in Prediction of Time Series", IEEE Transactions on fuzzy systems, Vol. 18, No. 2, pp 233-250, April 2010.

[3] Amit Mishra and Zaheeruddin, "design of hybrid fuzzy neural network for function approximation", J.Intelligent learning systems and applications, 2, pp97-109, May 2010.

[4] Antonios D. Niros and George E. Tsekouras, "A novel training algorithm for RBF neural network using a hybrid fuzzy clustering approach", fuzzy sets and systems, 163, pp 62-84, 2012.

[5] Sandeep Paul and Statish Kumar, "Subsethood product fuzzy neural inference system", IEEE Transactions on neural networks, Vol.13, No.3, pp578-599, May 2002.

[6] Sandeep Paul and Statish Kumar, "Subsethood based adaptive linguistic networks for pattern classification", IEEE Transactions on systems, man and cybernetics- part c: applications and reviews Vol.33, No.2, pp248-258, May 2003.

[7] Sandeep Paul and Statish Kumar, "fuzzy neural inference system using mutual subsethood products with applications in medical diagnosis and control", IEEE international fuzzy systems conference, pp 728-731, 2001.

[8] C.Shunmuga Velayutham and Statish Kumar, "Some applications of an Asymmetric Subsethood Product Fuzzy Neural Inference System", IEEE international conference on fuzzy systems, pp 202- 207, 2003.

[9] C.Shunmuga Velayutham and Statish Kumar, "Asymmetric Subsethood-Product Fuzzy Neural Inference System (ASuPFuNIS)", IEEE Transactions on neural networks, Vol. 16, No. 1, pp 160 – 174, January 2005.

[10] Hesham A.Hefny, "Comment on: Distinguishability quantification of fuzzy sets", Information Sciences (177), pp 4832-4839, February 2007.

[11] Nelly S. Amer and Hesham A.Hefny, "Analytical formulas for similarity, possibility and distinguishability measures of Cauchy type fuzzy sets with comparison to Gaussian fuzzy sets", Proceedings of the 7th IEEE International Conference on Intelligent Computing and Information Systems (ICICIS15), vol. 3, pp. 22-27, 2015.

[12] Shifei Ding and Fengxing Jin, "A Novel Fuzzy Likelihood Measure Algorithm", IEEE International conference on computer Sciences and software Engineering, pp 945-948, 2008.

[13] Shang- Ming Zhou and John Q.Gan, "Low –level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling", Fuzzy Sets and Systems 159, pp 3091-3131, June 2008.

Survey on Land Change Modeling

Khalid A. Eldrandaly[a], Mamdouh M. Abdeen[b], Safa A. Abdelkareem[c]

Abstract

The presented paper covers the subject of land change modeling, including the current approaches, applications, and the future prospects. A lot of classifications schemes do exist for the land change modeling approaches, many of which don't cover the massive variety of the land change models. Land change models requires to be organized and grouped to ease the procedure of selecting the most suitable one given a certain application or to satisfy a specific purpose. Applications of land change models and the preferred model for each application form a critical step included within the modeling land change process. The more precise the model fits a certain application including all its requirements and specifications the higher the accuracy of the simulated results, and more related to real world they will turn out to be. Although the last decade showed a great development not only in the field of land change modeling, but also in the land change science as a whole, still a number of problems do exists and needs to be addressed in order to move the land change modeling to the next level. Despite of the existing problems, chances for developments in the future are present. Developments such as the integration between the different approaches of land change modeling must be further studied to benefit from the advantages of different approaches and to limit drawbacks that accompany each approach. Some other future development ideas described in the future prospects section.

Key words: Land use/cover change, Land change modeling, LULCC applications.

1. Introduction

Land use land cover change(LULCC) presents the way human being change and modify the surface of Earth[1]. Land change modeling approaches form the primary foundation for creating a model, a model that will not only be concerned with the study of changes in land use and land cover, but also will identify the causes and consequences of such a change [2]. Prediction of future changes and simulation of different scenarios based on different policies can also be estimated using land change models. Such capabilities form a great interest in different disciplines [3], [4]. This paper aims to, discuss and review the current practices of land change modeling approaches; to identify a set of models and some examples of its applications, and to propose ideas for improving land change modeling. This paper is organized to discuss some of the different classifications of land change modeling approaches, pick out the most appropriate classification that does handle most, if not all models of land change, discuss the conceptual basis that convey the idea of each approach and the models built upon it, also to talk about the cons and pros of each approach, and to demonstrate each approach separately and in comparison with each other. Also to demonstrate different applications and models that may be used to study those applications. Based on the discussion of different approaches, models, and applications opportunities to improve the future of land change models will be shown in the last sections of the paper.

[a, c] Information Systems Department, Faculty of Computers and Informatics, Zagazig University, Al-Sharqiyah, Egypt
[b] Remote Sensing and Space Sciences, 23 Joseph Broz Tito St., El-Nozha El-Gedida, Cairo, Egypt

Section 2 will discuss the current state of land change modeling approaches. Although many approaches do exist it would be helpful if each approach can be classified under a particular category. The current practices section discusses a classification of different approaches that can include almost all land change modeling approaches under one of its categories. Section 3discusses the applications that use pattern based models to model past changes and predict future ones. . Section 4 of this survey paper focus on the future prospects proposing ideas for possible advancements that can help improve land change modeling. Finally section 5 concludes this paper.

2. Current practices:

Land change modeling approaches can be sorted according to different features. Although plenty approaches do exist[1], [2], [5]–[7]. Some researchers focused on a specific set of approaches such as spatially explicit modeling approaches [8], [9] or approaches for descriptive models (All, 2016). Others described a much wider range of approaches [1], [4], but still a much general classification that all or at least most of the approaches can be categorized under one of its classes. The classification proposed in [10], [11] suggested different approaches for land change modeling (Figure. 1). The classification covered a wide range if not all of modeling approaches and maintained a common ground to distinguish between modeling approaches while keeping in mind the theoretical and practical structure of each approach. Anyone who is interested in the theoretical and empirical foundation and much more detail of any given modeling approach can refer to previous reviews i.e.[10], [12]

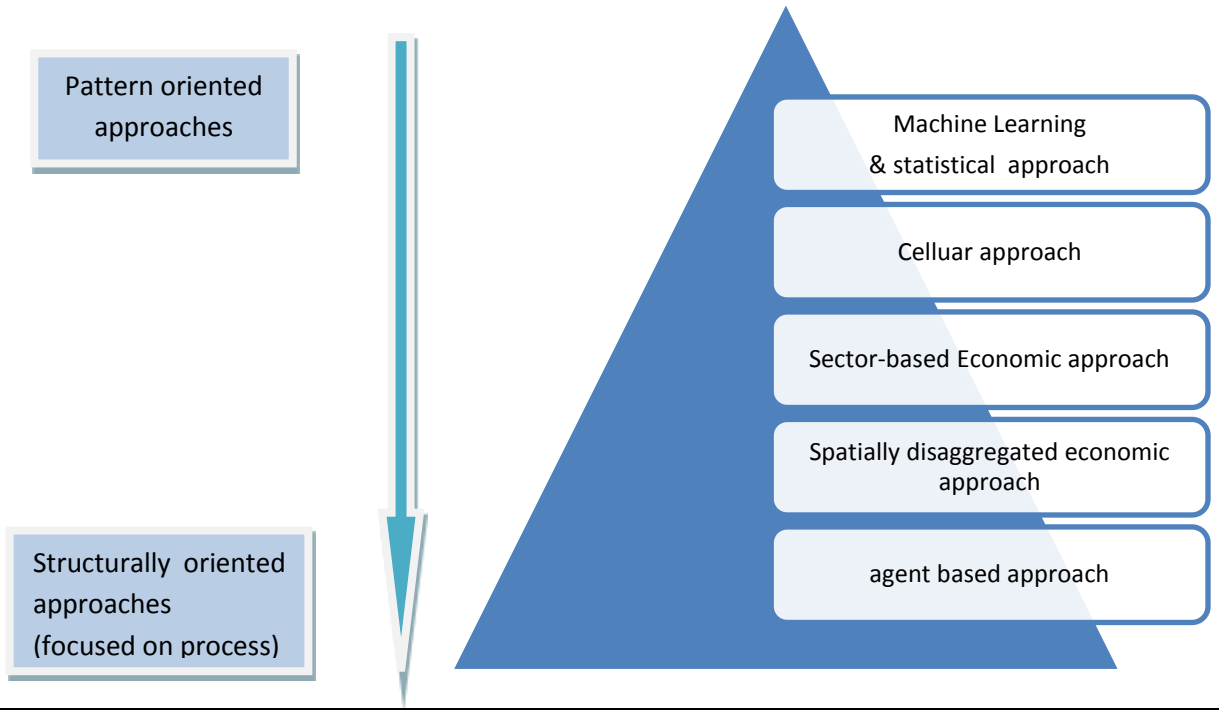


Figure.1. LCM Approaches: current practices.

2.1. Machine learning and statistical approach

This approach is considered to be a pattern based approach, and it has been using some kind of statistical model along with algorithms (i.e., Classification and regression trees, logistic regression, and artificial neural networks[7], [13]) to demonstrate relationships between changes of land use or cover and leading variables of the locations where they are likely to take place[4], [10], [12]. Machine learning is more appropriate whenever data related to patterns can be obtained while theory required for process can't. Whenever short-term views and predictions of the near future are in need machine learning approaches proven to be useful. In event of huge data or larger quantity of data sets the ability of the algorithms contained within this approach have proven to perform highly. This particularly means that the algorithm can run with limited computational requirements[4]. The algorithms are particularly used to project and generalize relationships found between inputs and outputs[10].

2.2 Cellular approach

The concept behind cellular automata is that each cell can carry a specific current state and can be moved to a next state which can be calculated using the current state, some transition rules and the neighbor cells also affect the process[4]. The bases of this method are shaped merging three components. The three components are suitability maps which indicate the degree in which a specific land can be used for a particular purpose[14], the second component focus on the impress of neighborhood on the land units used to mimic the study area, and the third component include information about the quantity of expected change of land[10].

The latter component enhances the power of the approach to predict and study the changes that are more likely to happen in the future[15]". The similarity between data structure in cellular models and land cover data abstracted from remote sensing makes it simple to start processing. Despite the simplicity of the three basic components forming cellular method seem to be, the format of models starts to develop a form of complicity when replacing the regularly formed pixels with spatial units. The selection of cellular models is not necessary because of the way it fits a specific case, sometimes one can choose to formulate the case study using cellular models due to their simplicity and well defined structure [10].

2.3 Econometric approaches

The third approach depends on the economic sector. Economic models connect outcomes that are generated from microeconomics actions that define the interaction between demand and supply to produce land use patterns[4]. The economic models determine the relationships between available and desired states which produces a link between prices and land use patterns. One of the features that characterize economic models its capability to explicitly determine how human being decisions affects land use, this quality is not contained within different approaches like the statistical or machine learning approaches[4]. this is particularly not an easy task taking in consideration the complexity of human

decisions and thinking mechanism[4]. The following two subsections and figure 2 define two different types of economic models. The sector based models and the spatially disaggregated models [4], [10]. The sector based models which can work on aggregated scale or spatially disaggregated models which operated at disaggregated scale.

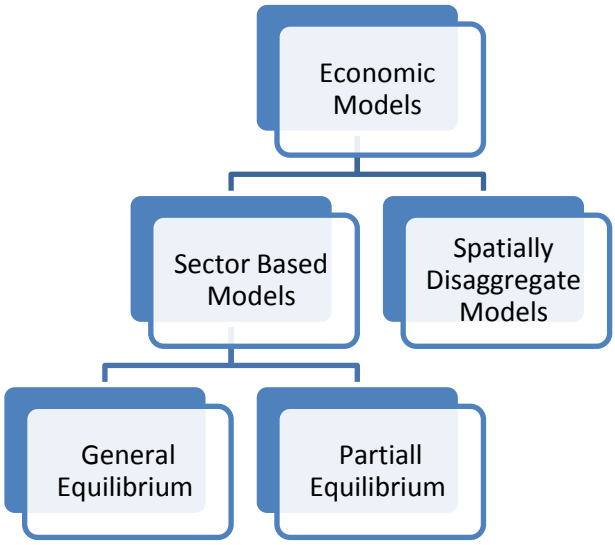


Figure.2: Economic models.

2.3.1Sector-Based Models

The sector based economic approach makes a great use of equilibrium structural models, whether general or partial, this is for the representation of demand and supply processes included in land system and defined by economic sectors which are based on trade and economic activity. Models of this approach basically concentrated on trade activities which can be detected in regions or sectors along with inputs and outputs, this is mainly to define demand and need for a specific land type. As mentioned before, those models are mostly structural which means they represent supply and demand explicitly as a contributing factor to market equilibrium. The scope of an economic system which is represented by sector-based model differentiates between models: partial equilibrium and general equilibrium models differ in their representation of the economic system. General equilibrium represents interactions and feedbacks between several sectors along with the global economy. While partial equilibrium models cares more about the details that describe a specific sector as a closed system, which means no linkages between the model and the rest of the outer economic.

Partial equilibrium models also determine prices, productions, and shares of land within a specific geographic area such as a country, afterwards uses the same land in different sectors. In addition, partial equilibrium models assume that the economic conditions in the rest of the world are fixed.

2.3.2 Spatially Disaggregated Economic Models

Spatially disaggregated economic models are thought to be structural models either fully structured or reduced structural. Reduced structural model mechanisms can be only implicitly illustrated through model specification and chosen variables. Models of this approach define the relationships that effect on the spatial equilibrium in land systems. One of the aims of this approach is to help investigators to fully interpret and acknowledge land-use and land-cover changes as an outcome of decisions made individually, this is possible because this approach models is designed based on the macroeconomic theory. The theory of microeconomics and the approaches built on its conceptual bases usually used to evaluate the implications of variables included in the model.

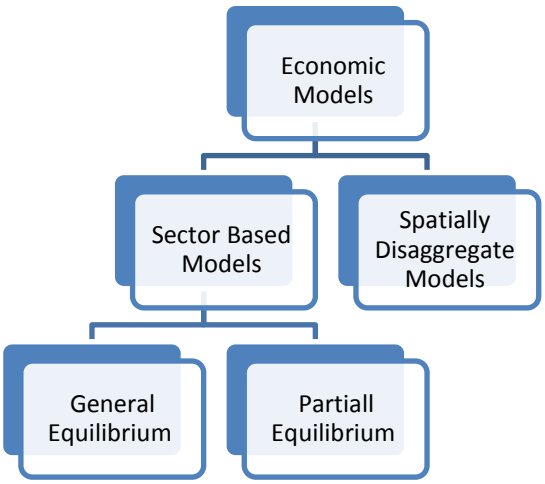


Figure.2: Economic models

2.4 Agent-Based Approach

Agent based models represent systems that are organized by a group of actors that interact with each other, these systems basically referred as multi-agent[4], [10]. When it comes to land change we can call land owners, farmers or any decision making figure as agents. Agents can be defined as any actor that contribute to the operation of decision making or the actor that perform a set of actions which has an effect on land use or land cover patterns[10]. Agents may be conceived as distinct objects that are accompanied by attributes and behaviors. Agents are associated with attributes that may be discretely or continuously measured. Interactions may occur between agents or between agents and surrounding environment this is particularly helpful in the process of collecting information or taking actions that affects their settings[2]. Although being spatially explicit is not always true for agent-based models, those involved in researching land change usually are. This means that agents and any associated actions are pointed at specific locations on Earth's surface. Agent based models are also considered to be structural models that use explicit processes to represent land change[10].

2.5 Hybrid Approaches

This section is dedicated to illustrate how a mix of the previously stated approaches can be used to address land change in the form of changed patterns and processes. Hybrid approaches are responsible for integrating theories and data for different environments[2]. Different theories, including economics and geography and many other contribute in the process of interpreting land change. Most theories cannot fully explain efficiently the complexity that is involved in land use decision making. The main concept that the hybrid modeling approaches are based on is its capability to merge various conceptual frames and theories while representing a system. The strengths of hybrid modeling approaches are derived from the individual strengths of each approach, also it minimize the drawbacks that may happen to each approach alone when they are combined[16].

2.6 Comparison of land change modeling approaches

Although the comparison of different models and modeling approaches may be difficult because of the different fundamental bases[1]. The following sub-sections include some aspects to compare the main modeling approaches stated above.

2.6.1 Pattern based vs. Process based

When describing the arrangement of the approaches, whether the approach focused on patterns or based on processes, the most patterns based approaches are the Machine Learning and Statistical and cellular approaches, while the Spatially Disaggregated Economic Models, Sector-Based Economic Models, and the Agent Based Models are more process focused as shown in figure. 1.

2.6.2 Likely outcome type

Each approach tented to be used for a certain outcome type. Machine Learning and Statistical approach most likely used to produce land cover. While Spatially Disaggregated Economic Models and Sector-Based Economic Models are usually used for land use. Both the land use and land cover have been produced from cellular approaches and Agent Based Models.

2.6.3 Data requirements

Usual input data requirements differ from one approach to the other, for instance the Machine Learning and Statistical approach requires land cover maps that were taken across two or more different time slots and some maps for single or multiple variables the will be used in the prediction process as inputs. While cellular approaches may require the same data inputs as machine learning approach the only difference it can only require a single map of land cover at a certain time point. The third approach discussed above (i.e. Spatially Disaggregated Economic Models) requires only data not maps like the previous two approaches about land use or land cover at one or more time points. Also takes economic and biophysical variables that greatly affect the process of demand and supply for land. Sector-Based Economic Models need the following inputs, the economic variables that effects on the aggregate demand and supply which include commodity price and trade value at regional or county scale. The final approach mentioned above which is the agent-based requires data for describing the agent characteristics, whether quantity or quality data involved in decision processes. Also requires data on land use or land cover at one or more time points.

2.6.4 Preferred uses

As shown in figure (4) Machine Learning and Statistical approach commonly used for producing predictions by recruiting with past patterns, while Cellular approaches recommended to be used for predicting land cover patterns, also for evaluation spatial changes without using market feedbacks. Spatially Disaggregated Economic Models recommended for recognizing the effect of the key variables on land change outcomes, this is true if the model was a reduced form model. While structural models of this approach used to simulate effects of changes of policy on the outcomes of the land market. Sector-Based Economic Models used to predict the aggregate land change which happened under a various set of changes built upon market which may affect the demand and supply. Agent-based requires data explores processes of land change usually under certain set of conditions. Also explores the effect of out changes in a system where they haven't occurred. Exploring future scenarios where past patterns proven to be insufficient and cannot refer to future patterns

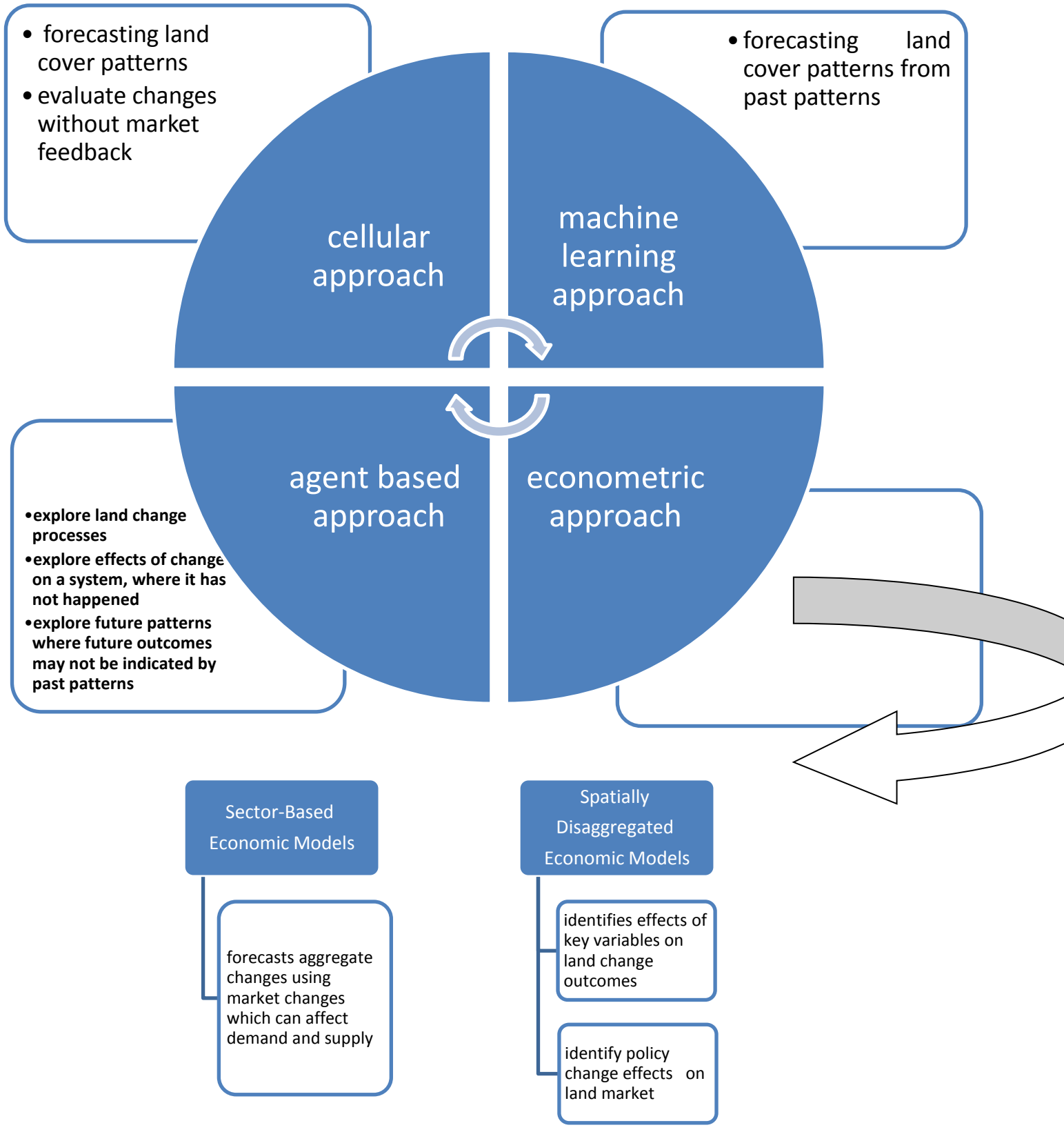

• forecasting land cover patterns
• evaluate changes without market feedback
cellular approach
machine learning approach
• forecasting land cover patterns from past patterns
agent based approach
econometric approach
•explore land change processes
•explore effects of change on a system, where it has not happened
•explore future patterns where future outcomes may not be indicated by past patterns
Sector-Based Economic Models
forecasts aggregate changes using market changes which can affect demand and supply
Spatially Disaggregated Economic Models
identifies effects of key variables on land change outcomes
identify policy change effects on land market

Figure.3. Recommended uses for each modeling approach

Figure (4) describes a hierarchy that the coming subsections will illustrate. Starting with the modeling approach or model methodology category, a subset of selected models was chosen based on how common they are or based on previous literature and has been reviewed by a number of technical experts in the field of land change science[18]. This is only a sample of models from each category farther models do exist, but they are not explained in this paper. This paper is more focused on the pattern based approaches which are the machine learning and statistical approach and the cellular approach this is why the models application's of those approaches are described in details while the other approaches are mentioned with some references for more information.

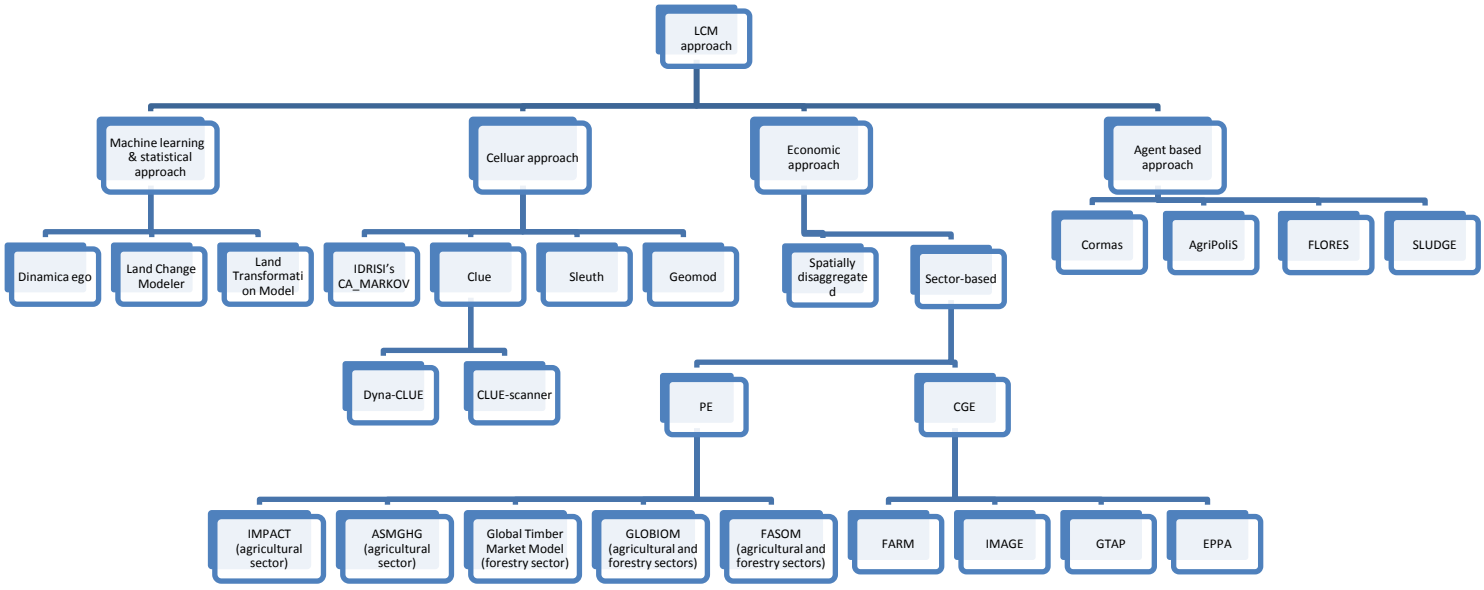


Figure.4. Hierarchical diagram to classify models used for the applications section

3. Applications of land change modeling approaches

The following figure describes a hierarchy that the coming subsections will illustrate. Starting with the modeling approach or model methodology category, a subset of selected models was chosen based on how common they are or based on previous literature. also a large number of the selected models where mentioned in the "Advancing Land Change Modeling: Opportunities and Research Requirements" book which was written by "Committee on Needs and Research Requirements for Land Change Modeling" and has been reviewed by a number of technical experts in the field of land change science. This is only a sample of models from each category farther models do exist, but they are not explained in this paper.

During the search conducted using both Google scholar scientific search engine and ISI's web of science/knowledge using the name of the model as a keyword, a large number of articles were found to be related to the topic. For the sake of simplification and organization farther search criterion was proposed so that no articles before the year 2010 were reviewed in this paper, also all articles that are not written in English were neglected and only the most rated or top cited or most recent articles when the search was conducted are discussed.

This paper is more focused on the pattern based approaches which are the machine learning and statistical approach and the cellular approach this is why the models application's of those approaches are described in details while the other approaches are mentioned with some references for more information.

3.1 Applications of Machine learning and statistical approach

Model name	Model description
Dinamica ego	is a software written in C++ and java and it includes a set of algorithms that help perform all the operations concerned with spatial analysis[19].
Land Change Modeler	Land change modeler is considered to be Terrset's (previously idrisi's) integrated set of tools designed for facilitating the process of assessment and prediction of land changes[20].
Land Transformation Model	The LTM considered being a neural network model. [21],the LTM relates both the potential (dependent) variables with the independent variables and develop numerical framework to include both inputs and output and trains on the relationship between them unit acceptable fit is reached to simulate land use cover change [21], [22].

Table.2 machine learning and statistical approach models example

Model name	Application	Integration with other models	References

Dinamica ego	deforestation, urban dynamics, agriculture expansion, forest fires	-	[23]
	to study the role of protected areas in reducing carbon fluxes caused by deforestation and the model helped predicting the carbon emission	carbon bookkeeping model	[24]
	to study changes in oil palm plantation expansion	-	[25]
	to study the effect of urbanization and climate changes on future flooding	climate change model	[26]
	to predict and model changes and growth of urban land	-	[27]
	deforestation modeling & projecting the future deforestation	-	[28]
	Modeling, calibrating, and validating fire regimes	-	[29]
	to optimize the ecosystem services	-	[30]
	helped studying the expected effects of agricultural expansion and climate change and how they both will affect on soil erosion	-	[31]
Land Change Modeler	to study urbanization, deforestation, habitat modeling and much more	-	[32]
	To monitor land cover changes of the dry land forest landscape.	-	[33]
	the impacts of land use changes and how it affects the erosion risk.	-	[34]
	maintain protected & to minimize biodiversity loss	climate change model	[35]
	to analyze urban sprawl & to model urban growth	-	[36]
	the spatial analysis of land abandonment	-	[37]
	To model the land use and climate patterns and how they affect surface water quality	-	[38]
	spatial planning and to produce different land-use zoning scenarios	-	[39]
	Model to assess and evaluate global biodiversity loss.	-	[40]

	to spatially asses farming systems and how they affect deforestation, forest re-growth, and agriculture change	-	[41]
	Timber plantation and forest expansion from other land cover and the analysis of causes of change	-	[42].
	Prediction of future land use land cover	-	[43]
	Urban growth analysis and prediction	-	[44],[45]
	The prediction of transformation to certain land category in the future	-	[46]
Land Transformation Model	to study urban growth	-	[47]
	Urban prediction	-	[48], [49]
	predict vacant land	-	[50]
	how land use and water quality affect each other	hydrologic models	[21]

Table.3 machine learning and statistical approach applications

3.2 Applications of Cellular Approach

The following subsections will discuss some of the most used cellular based models to simulate land changes and to project future changes.

Model name	Model description
IDRISI's CA-MARKOV	The CA-Markov model is a model that is included in the IDRISI Selva software, and it is essentially used to simulate land cover changes and to predict the future patterns[51].
Clue	The clue model is concerned with studying the conversion of land use and how such changes happen and what effects will arise from such changes. The clue model can operate at international, continental, and regional scales based on the required application and study area. New versions of the CLUE model has been developed two of the latest versions used are: Dyna-CLUE and CLUE-Scanner [52].
Sleuth	Sleuth model is considered to be one of known and strongly tested examples of cellular automata models[4], [53]..Sleuth contains two coupled CA models: the land change model and the urban growth model to asses historical changes and predict future land use patterns[54].
Geomod	The Geomod model had two versions Geomod1 and Geomod2, and it was integrated into IDRISI. Geomod can be applied on continental, country, and local scale and it has been used to study and analyze land use land cover changes specially scenarios of deforestation as it simulates the spatial patterns of changes through time[55].

Table.4 Cellular Approach models example

Model name	**Application**	**Integration with other**	**References**

		models	
IDRISI's CA-MARKOV	to predict and simulate the future land use and land cover changes	-	[56]
	to simulating urban expansion/growth	-	[57]
	To simulate developments in future land use that will affect the future of wild ungulates	-	[58].
	to produce maps projecting spatial and temporal changes of wetland	-	[51]
	the study and assessment of the biodiversity	-	[59]
	assist in the process of land resources management	-	[60]
	To identify changes to land use and climate that may eventually affect the water resources	climate change model	[61]
Clue	to predict land use changes in the future & the simulation of land cover changes but at a regional scale & to study rapid changes in phenomenon such as agriculture, urban growth, and trade sector at regional scale & land use transitions & farmland abandonment, deforestation, and carbon sequestration	-	[32], [62]–[66]
Sleuth	to study urban growth or development and to project future scenarios	-	[53], [67]–[78]
	to help in the process of assessing flood hazards	-	[79]
Geomod	to study and assess deforestation and to predict future	-	[80]
	to predict how one land state change from one state to another	-	[81]
	to assess and predict the changes of urban expansion scenarios and how it relates to floods and droughts	-	[82]
	predicting deforestation in the future	-	[83]

Table.5 Cellular Approach applications

4. Future prospects & trends

The Current and future direction all is focused on land change science and advancements in how to better detect past changes and how to accurately predict the future changes as close to reality as possible. The next few points will discuss future research trends.

- A major topic in advancing land change modeling is concerned with the development of the land change models validation, evaluation, and calibration practices, as they help both the model accuracy and advancement also they increase the degree of how well accepted the model to the user. More research must be conducted to further improve calibration practices specially the manual calibration and expert knowledge needs more attention also the way calibration is conducted in applications must be more defined[11], [16], [92].
- A number of researchers tried to facilitate the process of selecting the appropriate model for a given research application by assessing various types of LUCC models. Although stating the pros and cons of each land change model made it a bit easier to select the suitable model for a certain research, it is still pretty much foggy. To help researcher's choice the most suitable model a certain process needs to be defined or a set of steps to ease and improve the quality of the application since it relays on the chosen model[12], [93], [94]. This also can be done with help of expert systems or by using a multi-criteria decision making method to help make the decision.
- Assessing a geographic phenomenon using land use/cover change techniques, Due to the limited amount of studies that focused on comparing various land use/cover change models to asses change [95] upcoming studies and research should be focused on such subject.
- Focusing on local scale land change models and coming up with different ideas to manage to transmit them to global scale can help advancing land change modeling more and better results may appear from such models on a wider scale [47]
- The amount of remote sensing datasets and images used to run some land change models requires new computing technique and new approach deal and manage large datasets especially high resolution ones that can be large in both size and number of files and will consume more processing time to create a more accurate result[50]. Big data techniques can be integrated to manage and run such large datasets, some articles written about using big data to higher the performance of a land change model but more studies needs to be performed in the same topic[47][96].
- Datasets and sources that are used as input for land change models also effects the quality of results and also requires attention to enhance the quality of such datasets in terms of scale and temporal aspects, also the way those datasets are processed and managed before entered to the land change model needs more attention spatially while classifying images and selecting the best algorithms that suites the purpose of the research[97].

5. Conclusion

This paper has reviewed the current practices, applications, and future prospects for land change modeling, although plenty approaches are available each one meets a certain set of objectives. The future of land change modeling is in the integration of its models with other techniques to better understood and simulate changes.

References

[1] J. Jokar Arsanjani, *Dynamic land use/cover change modelling*. 2012.

[2] M. J. . D. and A. . V. Peter H . Verburg , Paul P . Schot, "Land use change modelling : current practice and research priorities," *GeoJournal*, vol. 61, no. 4, pp. 309–324, 2004.

[3] C. Agarwal, G. L. Green, M. Grove, T. Evans, and C. Schweik, "A Review and Assessment of Land-Use Change Models : Dynamics of Space , Time , and Human Choice Chetan Agarwal," *Apollo Int. Mag. Art Antiq.*, vol. 1, pp. 812–855, 2002.

[4] J. Li and X. Yang, "Monitoring and Modeling of Global Changes: A Geomatics Perspective," pp. 265–279, 2015.

[5] E. Koomen, J. Stillwell, A. Bakema, and H. J. Scholten, *Modelling Land-Use Change: Progress and Applications*, vol. 90, no. 2. 2007.

[6] P. H. Verburg, P. P. Schot, M. J. Dijst, and A. Veldkamp, "Land use change modelling: Current practice and research priorities," *GeoJournal*, vol. 61, no. 4, pp. 309–324, 2004.

[7] J. Li and X. Yang, *Jonathan Li Xiaojun Yang*. .

[8] D. G. Brown, R. Walker, S. Manson, and K. Seto, "Land Change Science," vol. 6, pp. 403–417, 2004.

[9] J. Mas, A. P. Vega, K. Clarke, and D. México, "Assessing ' spatially explicit ' land use / cover change models Centro de Investigaciones en Geografía Ambiental - Universidad Nacional Autónoma Department of Geography – University of California - Santa Barbara , USA," *Landsc. Ecol.*, pp. 508–513, 2010.

[10] D. G. Brown *et al.*, *Advancing Land Change Modeling*. 2014.

[11] D. G. Brown *et al.*, "Opportunities to improve impact, integration, and evaluation of land change models," *Curr. Opin. Environ. Sustain.*, vol. 5, no. 5, pp. 452–457, 2013.

[12] D. G. Brown, P. H. Verburg, R. G. Pontius, and M. D. Lange, "Opportunities to improve impact, integration, and evaluation of land change models," *Curr. Opin. Environ. Sustain.*, vol. 5, no. 5, pp. 452–457, 2013.

[13] J.-F. Mas, M. Kolb, M. Paegelow, M. C. Olmedo, and T. Houet, "Modelling Land use / cover changes : a comparison of conceptual approaches and softwares," *Environ. Model. Software*, vol. 51, pp. 94–111, 2014.

[14] J. Malczewski, "GIS-based land-use suitability analysis: A critical overview," *Prog. Plann.*, vol. 62, no. 1, pp. 3–65, 2004.

[15] Committee on Needs and Research Requirements for Land Change Modeling, Geographical Sciences Committee, Board on Earth Sciences and Resources, Division on Earth and Life Studies, and National Research Council, *Advancing Land Change Modeling: Opportunities and Research*. 2014.

[16] C. on N. and R. R. for L. C. M. G. S. C. B. on E. S. and R. D. on E. and L. S. N. R. Counci, *Advancing Land Change Modeling: Opportunities and Research Requirements*. .

[17] J. van V. and P. Verburg, "Land Use Modelling," 2015.

[18] C. on N. and R. R. for L. C. M. G. S. C. B. on E. S. and R. D. on E. and L. S. N. R. Council, *Advancing Land Change Modeling: Opportunities and Research Requirements*. 2014.

[19] T. GUO, "Dinamica_EGO_guidebook.pdf," Jan. 2015.

[20] L. C. Modeler, "Geospatial Monitoring and Modeling System Habitat and Biodiversity Modeler Ecosystem Services Modeler Earth Trends Modeler Climate Change Adaptation Modeler TerrSet Geospatial Monitoring and Modeling System," pp. 1–8.

[21] B. C. Pijanowski, M. M. Building, D. Hyndman, E. Lansing, and M. M. Building, "The application of the land transformation, groundwater flow and solute transport models for michigan's grand traverse bay watershed," pp. 1–12.

[22] B. C. PIJANOWSKI, K. T. ALEXANDRIDIS, and D. MULLUERE, "Modelling urbanization patterns in two diverse regions of the world," vol. 1, no. December, 2006.

[23] T. Centro, S. Remoto, U. Federal, and D. M. Gerais, "Modeling Environmental Dynamics with Dinamica EGO : Training Session in PORTUGAL How Dinamica EGO works ?"

[24] A. B. Soares-filho *et al.*, "Role of Brazilian Amazon protected areas in climate change mitigation Role of Brazilian Amazon change mitigation protected areas in climate," 2014.

[25] K. M. Carlson, L. M. Curran, D. Ratnasari, A. M. Pittman, and B. S. Soares-filho, "Committed carbon emissions , deforestation , and community land conversion from oil palm plantation expansion in West Kalimantan , Indonesia," vol. 109, no. 19, 2012.

[26] H. T. L. Huong and A. Pathirana, "Urbanization and climate change impacts on future urban flooding in Can Tho city , Vietnam," pp. 379–394, 2013.

[27] S. Berbero, A. Akın, and K. C. Clarke, "Landscape and Urban Planning Cellular automata modeling approaches to forecast urban growth for adana , Turkey : A comparative approach," vol. 153, pp. 11–27, 2016.

[28] F. Hajek, M. J. Ventresca, J. Scriven, and A. Castro, "Regime-building for REDD + : Evidence from a cluster of local initiatives in south-eastern Peru," *Environ. Sci. Policy*, vol. 14, no. 2, pp. 201–215, 2011.

[29] R. Almeida *et al.*, "Simulating fire regimes in the Amazon in response to climate change and deforestation Michael Coe , Hermann Rodrigues and Renato Assunção Published by : Wiley Stable URL : http://www.jstor.org/stable/23023102 Accessed : 09-08-2016 21 : 17 UTC Your use of ," vol. 21, no. 5, pp. 1573–1590, 2016.

[30] M. T. Coe *et al.*, "Deforestation and climate feedbacks threaten the ecological integrity of south – southeastern Amazonia," no. figure 1, 2013.

[31] E. E. Maeda, P. K. E. Pellikka, M. Siljander, and B. J. F. Clark, "Geomorphology Potential impacts of agricultural expansion and climate change on soil erosion in the Eastern Arc Mountains of Kenya," *Geomorphology*, vol. 123, no. 3–4, pp. 279–289, 2010.

[32] J. Mas, M. Kolb, M. Paegelow, M. Teresa, C. Olmedo, and T. Houet, "Environmental Modelling & Software Inductive pattern-based land use / cover change models : A comparison of four software packages," *Environ. Model. Softw.*, vol. 51, pp. 94–111, 2014.

[33] J. J. Schulz, L. Cayuela, C. Echeverria, and J. Salas, "Monitoring land cover change of the dryland forest landscape of Central Chile (1975 – 2008)," vol. 30, pp. 436–447, 2010.

[34] M. Leh, S. Bajwa, and I. Chaubey, "IMPACT OF LAND USE CHANGE ON EROSION RISK : AN INTEGRATED REMOTE SENSING , GEOGRAPHIC INFORMATION SYSTEM AND MODELING METHODOLOGY," 2011.

[35] F. V Faleiro, R. B. Machado, and R. D. Loyola, "Defining spatial conservation priorities in the face of land-use and climate change," *Biol. Conserv.*, vol. 158, pp. 248–257, 2013.

[36] M. G. Tewolde, 1, * and Pedro Cabral, and 2, "Urban Sprawl Analysis and Modeling in Asmara, Eritrea," *Remote Sens.*, pp. 2148–2165, 2011.

[37] G. I. Díaz, L. Nahuelhual, C. Echeverría, and S. Marín, "Landscape and Urban Planning Drivers of land abandonment in Southern Chile and implications for landscape planning," *Landsc. Urban Plan.*, vol. 99, no. 3–4, pp. 207–217, 2011.

[38] C. O. Wilson and Q. Weng, "Science of the Total Environment Simulating the impacts of future land use and climate changes on surface water quality in the Des Plaines River watershed ,

Chicago Metropolitan Statistical Area , Illinois," *Sci. Total Environ.*, vol. 409, no. 20, pp. 4387–4405, 2011.

[39] D. Geneletti, "Assessing the impact of alternative land-use zoning policies on future ecosystem services," *Environ. Impact Assess. Rev.*, vol. 40, pp. 25–35, 2013.

[40] A. Pérez-Vega, J. F. Mas, and A. Ligmann-Zielinska, "Comparing two approaches to land use/cover change modeling and their implications for the assessment of biodiversity loss in a deciduous tropical forest," *Environ. Model. Softw.*, vol. 29, no. 1, pp. 11–23, 2012.

[41] A. Carmona, L. Nahuelhual, C. Echeverría, and A. Báez, "Linking farming systems to landscape change : An empirical and spatially explicit study in southern Chile," *"Agriculture, Ecosyst. Environ.*, vol. 139, no. 1–2, pp. 40–50, 2010.

[42] L. Nahuelhual, A. Carmona, A. Lara, C. Echeverría, and M. E. González, "Landscape and Urban Planning Land-cover change to forest plantations : Proximate causes and implications for the landscape in south-central Chile," *Landsc. Urban Plan.*, vol. 107, no. 1, pp. 12–20, 2012.

[43] K. S. Kumar, P. U. Bhaskar, and K. Padmakumari, "APPLICATION OF LAND CHANGE MODELER FOR PREDICTION OF FUTURE LAND USE LAND COVER A CASE STUDY OF VIJAYAWADA CITY," pp. 2571–2581.

[44] Y. Megahed, "LAND COVER MAPPING ANALYSIS AND URBAN GROWTH MODELING USING REMOTE SENSING TECHNIQUES Case Study : Greater Cairo Region - Egypt Case Study : Greater Cairo Region - Egypt."

[45] R. B. Thapa and Y. Murayama, "Computers , Environment and Urban Systems Urban growth modeling of Kathmandu metropolitan region , Nepal," *Comput. Environ. Urban Syst.*, vol. 35, no. 1, pp. 25–34, 2011.

[46] M. I. Mahmoud, A. Duker, C. Conrad, M. Thiel, and H. S. Ahmad, "Analysis of settlement expansion and urban growth modelling using geoinformation for assessing potential impacts of urbanization on climate in Abuja City, Nigeria," *Remote Sens.*, vol. 8, no. 3, 2016.

[47] B. C. Pijanowski, A. Tayyebi, J. Doucette, B. K. Pekin, D. Braun, and J. Plourde, "A big data urban growth simulation at a national scale: Configuring the GIS and neural network based Land Transformation Model to run in a High Performance Computing (HPC) environment," *Environ. Model. Softw.*, vol. 51, pp. 250–268, 2014.

[48] S. Park, S. Jeon, S. Kim, and C. Choi, "Landscape and Urban Planning Prediction and comparison of urban growth by land suitability index mapping using GIS and RS in South Korea," *Landsc. Urban Plan.*, vol. 99, no. 2, pp. 104–114, 2011.

[49] U. Projection, "Land Transformation Model," pp. 11–20, 2000.

[50] G. Newman, J. Lee, and P. Berke, "Using the land transformation model to forecast vacant land," *J. Land Use Sci.*, vol. 4248, no. March, pp. 1–26, 2016.

[51] V. M. Bacani, A. Y. Sakamoto, H. Quénol, C. Vannier, and S. Corgne, "Markov chains–cellular automata modeling and multicriteria analysis of land cover change in the Lower Nhecolândia subregion of the Brazilian Pantanal wetland," *J. Appl. Remote Sens.*, vol. 10, no. 1, p. 016004, 2016.

[52] P. Verburg, "The CLUE model," no. Hands-on exercises, p. 53, 2010.

[53] X. Li and P. Gong, "Urban growth models: progress and perspective," *Sci. Bull.*, vol. 61, no. 21, pp. 1637–1650, 2016.

[54] K. C. Clarke, "Cellular Automata and Agent-Based Models," pp. 1217–1233.

[55] R. G. Pontius Jr. and H. Chen, "GEOMOD Modeling," *Clark Labs*, no. January 2006, pp. 1–44, 2006.

[56] H. Mahmoud and P. Divigalpitiya, "Modeling Future Land Use and Land-Cover Change in the Asyut Region Using Markov Chains and Cellular Automata," *Smart Sustain. Plan. Cities Reg.*, pp. 1–435, 2017.

[57] M. M. Aburas, Y. M. Ho, M. F. Ramli, and Z. H. Ash'aari, "Improving the capability of an integrated CA-Markov model to simulate spatio-temporal urban growth trends using an Analytical Hierarchy Process and Frequency Ratio," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 59, pp. 65–78, 2017.

[58] P. Acevedo, M. Á. Farfán, A. L. Márquez, M. Delibes-Mateos, R. Real, and J. M. Vargas, "Past, present and future of wild ungulates in relation to changes in land use," *Landsc. Ecol.*, vol. 26, no. 1, pp. 19–31, 2011.

[59] P. Mondal and J. Southworth, "Evaluation of conservation interventions using a cellular automata-Markov model," *For. Ecol. Manage.*, vol. 260, no. 10, pp. 1716–1725, 2010.

[60] S. S. Palmate, A. Pandey, and S. K. Mishra, "Modelling spatiotemporal land dynamics for a trans-boundary river basin using integrated Cellular Automata and Markov Chain approach," *Appl. Geogr.*, vol. 82, pp. 11–23, 2017.

[61] S. T. Y. Tong, Y. Sun, T. Ranatunga, J. He, and Y. J. Yang, "Predicting plausible impacts of sets of climate and land use change scenarios on water resources," *Appl. Geogr.*, vol. 32, no. 2, pp. 477–489, 2012.

[62] A. Tayyebi, B. K. Pekin, B. C. Pijanowski, D. James, J. S. Doucette, and D. Braun, "Hierarchical modeling of urban growth across the conterminous USA : developing meso-scale quantity drivers for the Land Transformation Model," *J. Land Use Sci.*, no. February 2015, pp. 37–41, 2012.

[63] L. D. A. Campania, S. Pindozzi, E. Cervelli, P. F. Recchi, and A. Capolupo, "Predicting land use change on a broad area: Dyna-CLUE model appication to the Litorale Domizio-Agro Aversano (Campania, South Italy)," vol. XLVIII, pp. 27–35, 2017.

[64] J. Fox, J. B. Vogler, O. L. Sen, T. W. Giambelluca, and A. D. Ziegler, "Simulating land-cover change in Montane mainland southeast Asia," *Environ. Manage.*, vol. 49, no. 5, pp. 968–979, 2012.

[65] G. Luo, C. Yin, X. Chen, W. Xu, and L. Lu, "Combining system dynamic model and CLUE-S model to improve land use scenario analyses at regional scale: A case study of Sangong watershed in Xinjiang, China," *Ecol. Complex.*, vol. 7, no. 2, pp. 198–207, 2010.

[66] E. E. Maeda, C. M. de Almeida, A. de Carvalho Ximenes A., A. R. Formaggio, Y. E. Shimabukuro, and P. Pellikka, "Dynamic modeling of forest conversion: Simulation of past and future scenarios of rural activities expansion in the fringes of the Xingu National Park, Brazilian Amazon," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 13, no. 3, pp. 435–446, 2011.

[67] F. Sunar and S. Berbero, "Urban change analysis and future growth of Istanbul," no. April, 2015.

[68] A. A. A. Alsharif and B. Pradhan, "Urban Sprawl Analysis of Tripoli Metropolitan City (Libya) Using Remote Sensing Data and Multivariate Logistic Regression Model," vol. 42, no. March, pp. 149–163, 2014.

[69] A. J. Terando, J. Costanza, C. Belyea, R. R. Dunn, A. McKerrow, and J. A. Collazo, "The southern megalopolis: Using the past to predict the future of urban sprawl in the Southeast U.S.," *PLoS One*, vol. 9, no. 7, 2014.