Evaluation of the packet loss in WSN using Deterministic Stochastic Petri Nets

Amel Berrachedi, Chourahbil Ben Mokhtar, and Malika Ioualalene

MOVEP, USTHB, Algiers, Algeria

Berrachedi_amel@yahoo.fr, chourahbil.benmokhtar@gmail.com, mioualalen@usthb.dz

Abstract. When developing critical and complex systems, the analysis and the performance evaluation of the designed system is a challenging task. Wireless Sensor Network (WSN) are examples of such systems. A WSN consists of a large amount of distributed autonomous nodes which monitor physical or environmental conditions. In this paper, we address the problem of how WSNs must be designed to fulfil system requirements and to have good performances in particular the packet loss ratio. To do so, we use an expressive kind of Petri Nets called Deterministic Stochastic Petri Nets. To show the applicability of the proposed model, one of routing techniques dedicated for WSNs is presented.

Keywords: WSNs, Deterministic Stochastic Petri Nets, Packet Loss.

1 Introduction

Since their appearance, the Wireless Sensor Networks (WSNs) increasingly invade the scientific and industrial communities. They are used in a wide range of applications, such as environmental monitoring, robotic exploration, military applications, medical systems and so on.

A WSN consist of a set of miniature, autonomous and multifunctional sensor nodes which are distributed on a capture zone to measure a physical magnitude or monitor an event. These nodes sense the information from their environment and relay them to a base station which is generally supposed powerful and away from the coverage area [6, 10].

As sensor nodes are primarily powered by irreplaceable and limited batteries, they

should work with a low energetic consumption. So, when designing such networks, the main requirement is to maximize the network lifetime that is overall influenced by protocols chosen for the deployed WSN. Among these protocols, the hierarchical ones attract the researchers' attention as they are so efficient in terms of energy.

A great number of research focuses on the energy constraints to the detriment of other factors, such as the packet loss. The hostile nature of WSNs environments generates nevertheless a lot of errors during the data packets routing and therefore these packets can be considerably lost over their transmission. Accordingly, if we manage to design a WSN with efficient power management without taking into account the significant number of losses, the network can't be reliable especially in critical applications (for example, security applications). Thus, the packet loss factor is essential when designing a WSN with good performances.

In order to prove that the designed network has good performances and the properties wanted by their designer, a phase of analysis and performance evaluation is necessary. The traditional approach is to evaluate the network protocols by simulators. However, a protocol can be considered as correct using the simulation, but many incidents in the real world have proved that this notion of correctness is insufficient. Indeed, such errors can result significant expenditures and jeopardize human beings life. The cause, that the simulation doesn't deal all possible cases of execution. In addition, a simple fault that occurs in a protocol can significantly reduce the WSN life and increase greatly the packet loss rate. The use of formal techniques becomes unavoidable [2]. Different formalisms have recently been used to design and analyze WSNs, and to evaluate their performances. Among these formalisms, Petri Nets (1962) [7] have many advantages; particularly, Deterministic and Stochastic Petri Nets (DSPN) could be the most appropriate for our specific problem. In fact, they are very expressive and they represent a widely used high-level formalism for modeling discrete-event systems where events may occur either without consuming time, after a deterministic time, or after an exponentially distributed time [9].

The aim of this work is to propose a WSN modeling approach for the evaluation of their performances, especially the packet loss constraints. It is necessary to consider this metric in order to minimize the number of lost data packets before implementing the routing protocols in the real world. In the case where the packet loss rate is high, the network designers must develop mechanisms to minimize these losses. The paper is organized as follows. Section 2 discusses the related work. We give in section 3, the formal definition of DSPNs, followed in Section 4 by an illustration of how to use it to model and evaluate the packet loss in a WSN. We present also some results for which we perform simulations using the TimeNET tool [11]; a graphical tool that allows the modeling, the analysis and the simulation of DSPNs. The section 5 provides some conclusions and possible improvements to the model.

2 Related Work

In the existing work, the packet loss concept wasn't deeply treated, in particular in the context of WSNs in which the major constraint is the energy consumption. In [4], the authors introduced the concepts of Non-Hierarchical Colored Petri Nets. This is done by means of a running example consisting of a simple communication protocol, in which a sender transfers a number of data packets to a receiver. The protocol uses sequence numbers, acknowledgements, and retransmissions to ensure that the data packets are delivered once and in the correct order. In WSNs, the retransmissions aren't permitted, because the node energy can significantly be decreased and nodes will be quickly exhausted. This is the reason why the packet losses are tolerated in WSNs.

In [8], a framework based on Stochastic Petri Nets for modeling and analysis the switching node QoS is proposed. The authors used a simple package mana-gement mechanism in WAN networks that rejects any packet arriving at the node when the queue is full. The aim is to anticipate the packet loss to reduce the arrival flow in this case. The lost packets are modeled by a drop transition and their number is evaluated by the ring average number of this transition.

In [1], we quantified the energy consumption on the Petri Net model assuming that packet loss doesn't occur. In reality, there are always packet losses. This depends on several factors, such as the hostile environments where the sensor nodes are deployed, the distance between the nodes, the data flow, etc. In this paper, we aim to consider the packet loss metric using the same case study in our previous work [1], where we modeled and evaluated the performances of a hierarchy protocol, based on the clustering technique. In addition, such as sensors' actions can be immediate as they can consume some time, it would be better to use a formalism which take into account these different kinds of operations, namely the Deterministic Stochastic Petri Nets.

3 Deterministic Stochastic Petri Nets

The Deterministic and Stochastic Petri nets (DSPNs), introduced by Ajmone Marsan and Chiola in [5], are a stochastic modeling formalism with graphical representation, which include both exponentially distributed and deterministic delays.

A DSPN is a 9-tuple (P, T, I, O, V, W, Π, D, M_0) [3], where:

- P is a finite set of places. $P = \{p_1, ..., p_n\}$;
- T is a finite set of transitions, disjoint from P, partitioned into three disjoint sets, T^I, T^E, and T^D, of immediate, exponential, and deterministic transitions respectively. $T = \{t_1, ..., t_m\}$;
- I is a set of the input arcs. $I \subseteq (P \times T)$;
- O is a set of the output arcs. $O \subseteq (T \times P)$;
- V is a set of the inhibitor arcs. $V \subseteq (P \times T)$, where $V \cap I = \emptyset$;
- W defines the weights of all arcs;
- Π is the priority function assigning a priority to each transition. $\Pi : T \rightarrow N^+$;
- D defines the firing times. $D : T \rightarrow 0 \cup R^+ \cup \Omega$, where R^+ is the set of positive real numbers and $\Omega = \{\lambda_1, ..., \lambda_l\}$ is the set of random variables with a given distribution;
- M_0 is the initial marking.

Graphically, the DSPNs have the same graphical notation of places and arcs in traditional Petri nets. However, the immediate transitions drawn as thin bars fire without delay, the exponential transitions drawn as empty bars re after an exponentially distributed delay, whereas the deterministic transitions drawn as black bars re after a constant delay.

Remark 1:

- The immediate transitions have a high firing priority relative to the timed ones. The default priority is 1 and the higher numbers mean the higher priorities.
- The random switches associate probability distributions to subsets of conflicting immediate transitions. So, for example, if the immediate transitions t_i and t_j are the only immediate transitions enabled, the probability of firing the transition t_i is given by $\frac{D(t_i)}{D(t_i) + D(t_j)}$
- The delay of an exponential distribution is given by $\frac{1}{\lambda_i}$, where $\lambda_i \in \Omega$ is the rate parameter of the distribution.

4 Case study: clustering technique

In a WSN based on a hierarchical topology, the sensor nodes are organized into clusters (see the figure 1). Each cluster is managed by a single node called Cluster

Head (CH). The remaining nodes are called members. During a time interval called round, each member picks up information from its environment and sends them to its CH. Used as gateway to reach the base station; each CH aggregates the received data and sends the final result to the base station which is assumed far from the coverage area. This means that only CHs manage the clusters, aggregate the data and transmit over long distances. The first technique used for hierarchical protocols is called static clustering, where the roles of nodes are pre-determined; a member will not become a CH and a CH will not become a member after deployment. In addition, the clusters have approximatively the same size and the sensor nodes are uniformly distributed. To reduce the complexity of our model, we take an example of a cluster of three nodes.

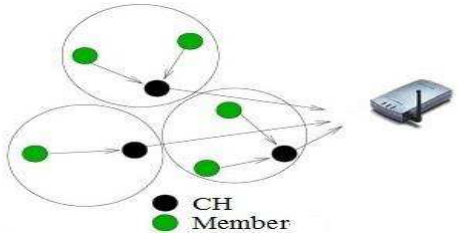

Fig. 1. An example of a hierarchical WSN.

As operations that the sensor nodes perform, can last either a predefined time (like the round which is fixed) or a random time (like emission/reception operations), the use of the Deterministic Stochastic Petri Nets is necessary. In fact, the predefined and random times are modeled by deterministic and exponential transitions respectively. Furthermore, priorities can be assigned to each transition. For example, a CH has to aggregate data even if it doesn't receive a packet. In fact, the latter can be wasted, so, the aggregation must have higher priority than the reception of members' data. Once the DSPN model is implemented, the properties analysis can be launched and the performances measures can be evaluated.

The idea is to model the sent packets, by a place, and the received packets by another place. These two places act as counters that will calculate the number of the sent packets and the number of the received packets. We calculate later the Packet Loss Ratio that is given by:

$$PLR = 1 - \frac{\text{Re}\,ceived\ Packets\ Number}{Sent\ Packets\ Number}$$

4.1 DSPN Model for the static clustering

The figure 2 shows a DSPN model associated with a cluster of size 3, i.e., the cluster consists of three sensor nodes; one CH and two members. The definitions of places and transitions are given in the tables 1 and 2 respectively.

Table 1. Definitions of the model places.

Place	Description
Init	Initial state
MBR_i	The member i is ready to sense information
$ReadyS_i$	The member i is ready to send information to his CH
$TransitD_i$	The data is being transmitted
Data	The members' data have probably been received and are awaiting aggregation
aggrattempt	Aggregation clock; its reset causes the end of the aggregation
SentD	Sent packets counter
AggrD	Received packets counter

Table 2. Definitions of the model transitions.

Transition	Description	Firing Distribution	Delay	Priority
StartRound	The members start their tasks	Immediate	0	4
$Sense_i$	The member i senses the information	Exponential	SenRate	1
S_iCH	The member i sends its own data to its CH		SRate	1
RD_i	The CH receives data of member i		RRate	1
Aggr	The CH aggregates the received data	Deterministic	ADelay	2
S-BS	The CH sends the final result to the base station	Deterministic	SBSDelay	3

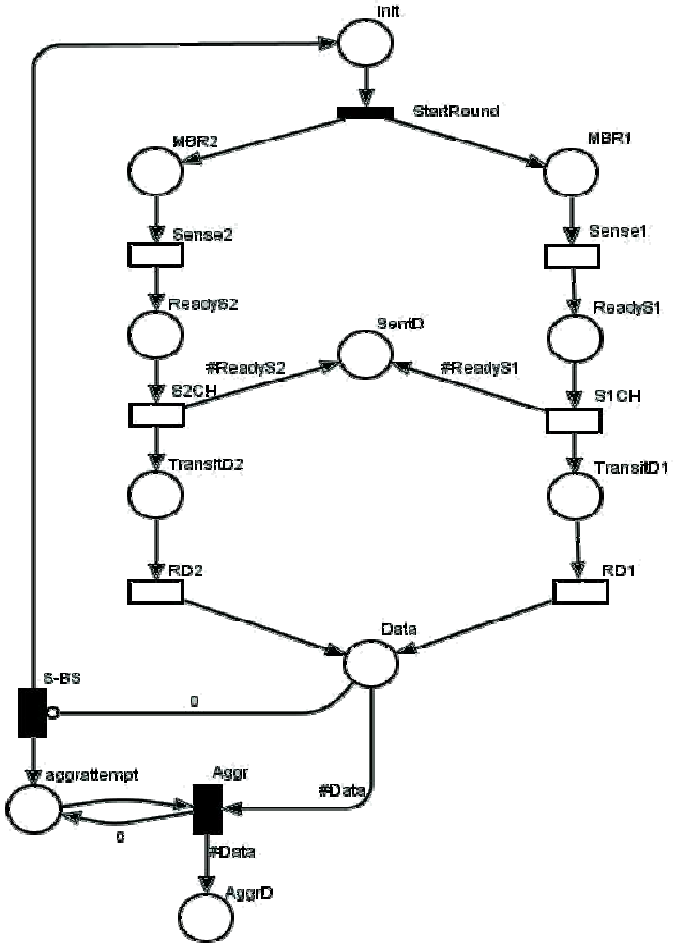

Fig. 2. The DSPN Model of a cluster with size 3.

The beginning of the round is determined by the initial state *Init*. As the immediate transition *StartRound* has the higher priority, so it is enabled and red without wasting time, because it doesn't last any time as the action performed by the sensor node doesn't admit any delay. The two members are ready (place MBR_i) to sense data from their environment. Then, they start sensing by firing the transition $Sense_i$ until a time $\frac{1}{SenRate_i}$ to send data. The transition S_iCH is enabled and can be red according to its

associated rate $SRate_i$. We will put a token in the place $TransitD_i$ and another in the place *SentD*. The latter acts as a counter of messages sent by the two members. So, it will keep track of the number of sent packets from the members to their CH.

The RD_i transition can be fired as it can't be. It depends on the sent packets of member i, if it hasn't been lost in transit. Therefore, the place *Data* may contain 2 tokens, as it can contain only one, and also, it can be empty. To avoid the network limit, the CH will not wait an infinite time. So, it has to launch the aggregating operation after the time associated to transition *Aggr* has passed. In addition, as it has a higher priority than other exponential transitions, it will be fired. The *Aggr* transition firing occurs when there is at least one token in the place *Data*. The weight *#Data* means that we will remove all the tokens of the place *Data*. During aggregation operation, the CH increases another counter *AggrD* which calculates the number of the received packets. If there is no token in the place Data, when the transition *S-BS* is enabled, it should be red using an arc inhibitor from the place *Data* to this transition. Finally, the nodes will be reset, and so, this is the end of the round. The same algorithm will be repeated for each round.

4.2 Packet loss evaluation

As the two places *SentD* and *AggrD* admit an infinite number of tokens, the model is not bounded. Then, the steady state is not reached. Consequently, we can't launch the numerical analysis of the stationary behavior. Timenet provides a transient numerical analysis for DSPNs by choosing an execution delay (1000 seconds in our case). The simulation component of TimeNET can perform the transient as well as the stationary evaluation of the Stochastic Petri Nets in continuous time, without the restriction of not more than one enabled transition with non-exponentially distributed ring time in each marking (see the figure 3).

As is illustrated in the figure 3, there are two types of measures: definition measures and performance ones [11].

- The definitions measures can be created by using the button *Def*, and they represent any expression that may be as an input value, like the ring rates which are predefined. For example, we create the definition measure $SenDelay = \frac{1}{SenRate}$ that should be adjusted in order to get the best results evaluation.

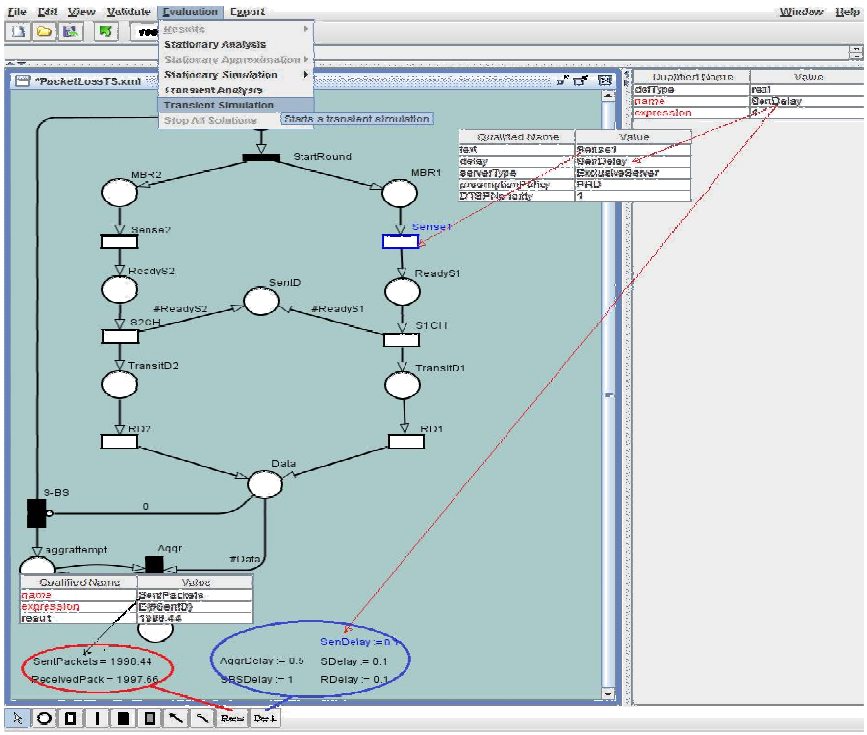

Fig. 3. DSPN Model of static clustering on Timenet.

Remark 2: The transition firing times are specified as delays for all transition types. The firing rates of the exponential transitions have to be transformed into delays into delays by taking their reciprocal value:

$$Delay_i = \frac{1}{\lambda_i}$$

The performances measures define what is computed during an analysis. A typical value would be the average number of tokens in a place. For example, the average number of the place *SentD* defines the sent packets number. We note: *E{#SentD}*. In addition, the average number of the place *AggrD* de nes the received packets number. In this case, we note: *E{#AggrD}*. Then, the Packet Loss Ratio is given by:

$$PLR = 1 - \frac{\mathrm{E}\{\#\,\mathrm{AggrD}\}}{\mathrm{E}\{\#\,\mathrm{SentD}\}}$$

Assuming that the length of a round is equal to 1 second, the delay associated to the transition *S-BS* is equal to 1 second. In other words, after a second has elapsed, the CH should send the final result to the base station, and the nodes must return to the initial state. This is why we associate a higher priority for this transition than other non-immediate transitions.

We set the aggregation time to 0.5 second. We assume that the time for each elementary action ($Sense_1$, $Sense_2$, etc.) is 0.1 second. We launch a transient simulation with an interval of 1000 seconds. The results and the associated graphs are given in 4 and 5. Both graphs show the increase of the number of the sent packets and the received packets over the time. They also show that the number of the received packets is less than the number of the sent packets.
The average of the packet loss rate is equal to 0.00185. It is insignificant. In our example, it only depends on the choice of ring rates that must be adjusted to improve the performances. The chosen routing protocol, which is based on the static clustering, is also simple. So, we have to make our model appropriate for other kinds of protocols. Furthermore, it would be better to consider other factors such as, other kinds of topology and different application domains. We must also think about the factors that increase the packet loss rate and included them in the proposed approach. For example, the distances, the impact of the environment and so on.

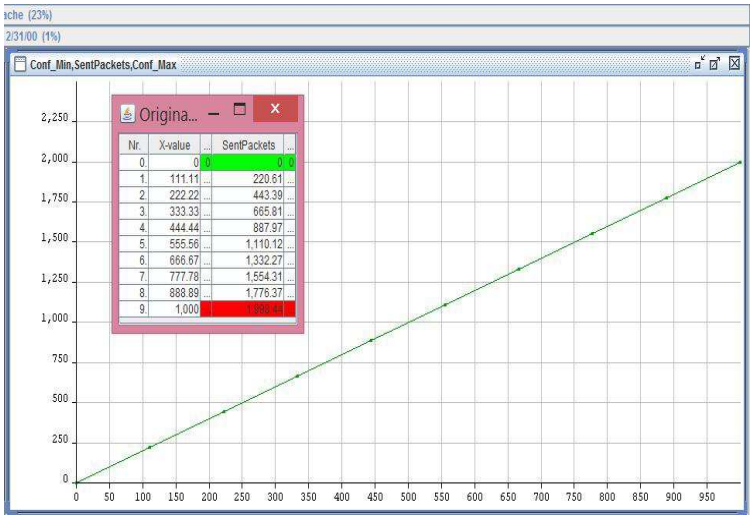

Fig. 4. The sent packets during 1000 seconds.

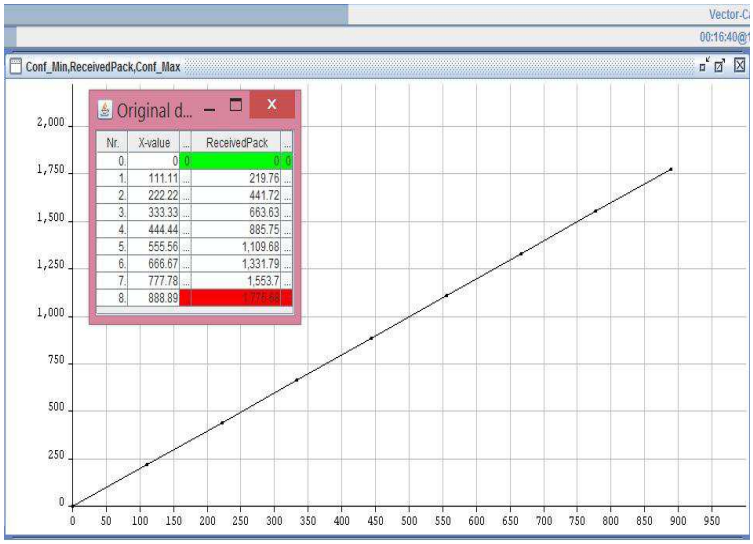

Fig. 5. The received packets during 1000 seconds.

5 Conclusion

Nowadays, Wireless Sensor Networks are a very active research topic. Based on the fact that the energy is the most important constraint, the researchers are more interested in the power consumption. However, it would be interesting to consider other constraints such as the packet loss which was been evaluated by few studies.

In this work, we evaluated the packet loss using formalisms based on the Petri Nets. Our choice fell on the Deterministic Stochastic Petri Nets because they seem more appropriate to our problematic.

In order to illustrate the behavior of the proposed DSPN model and evaluated the packet loss, we modeled one of the commonly proposed solutions for WSN routing namely the static clustering. The results showed that there are no high packet losses. In fact, we didn't take into account the factors that help to increase the packet loss.

In future work, we aim to find these factors that influence the packet loss metric. Important work remains to be done, to provide a formal framework for better properties verification and performance evaluation of the sensor network technology.

References

1. A. Berrachedi and M. Ioualalen. Modlisation de la notion d'énergie dans les réseaux de capteurs l'aide des réseaux de petri. In The 1st International Conference on Advanced Communication and Information Systems, pages 178-185, Batna University of Algeria, December 2012.
2. Jackson Francomme, Karen Godary, and Thierry Val. Validation formelle d'un mcanisme de synchronisation pour réseaux sans fil. In CFIP'2009, October 2009.
3. Mario Giacobini. Applications of Evolutionary Computing. EvoWorkshops 2007: EvoCOMNET, EvoFIN, EvoIASP, EvoINTERACTION, EvoMUSART, EvoSTOC, and EvoTransLog, Valencia, Spain, 2 April 2007.
4. K. Jensen and L.M. Kristensen. Chapter 2: Non-hierarchical Coloured Petri Nets, pages 13-41. 2009.
5. M. Ajmone Marsan. Stochastic petri nets: An elementary introduction. In In Advances in Petri Nets, pages 1-29. Springer, 1989.
6. Catello Di Martino. Resiliency assessment of wireless sensor networks: a holistic approach. PhD thesis, 'Federico II' University of Naples, Italy, December 2009.
7. James L. Peterson. Petri nets. ACM Computing Surveys, 9(3):223-252, 1977.
8. Tolotriniaina Mirado RAJAONARISON. Use of petri network for the study of switching node in a network wan. Master thesis, UNIVERSITE D'ANTANANARIVO, 2010.
9. Christian Rohr. Simulative csl model checking of stochastic petri nets in idd-mc.
10. Bashir Yahya, Jalel Ben-Othman, Lynda Mokdad, and Serigne Diagne. Perfor-mance evaluation of a medium access control protocol for wireless sensor networks using petri nets. In HET-NETs'2010, pages 335-354, 2010.
11. Armin Zimmermann and Michael Knoke. TimeNET 4.0: A Software Tool for the Performability Evaluation with Stochastic and Colored Petri Nets. Real-Time Sys-tems and Robotics Group, Technische Universit at Berlin, faculty of eecs technical report 2007-13 edition, August 2007.

Bat algorithm for overlapping community detection

Messaoudi Imene[1] and Kamel Nadjet[2]

[1] University of Sciences and Technology Houari Boumediene, BP 32 EL ALIA 16111 BAB EZZOUAR ALGIERS

messimene@yahoo.com,
WWW home page: http://www.usthb.dz/fei/
[2] University Setif 1, Faculty of Sciences, Departement of Computer Science. UFAS1. Setif, Algeria.

nkamel@univ-setif.dz

Abstract. In social network, a group of elements sharing common interests is called community. To know the structure of these communities, many works have been proposed with different techniques; we can cite label propagation, clique percolation, local expansion, etc. This structure is complex where communities overlap every time. In this paper we use bat algorithm to discover overlapping communities. Bat algorithm is a novel metaheuristic which is characterized by the echolocation behavior of bats. The algorithm we propose in this paper is based on the links of the network. The objective function evaluates the link density which is convenient for overlapping communities. Experiment on real networks show that the communities discovering with our approach have a higher density.

Keywords: overlapping community, link community, bat algorithm

1 Introduction

A social network is a collection of social objects (individuals, organizations) and relationship between these objects. The vision of a social structure as a network facilitates the understanding and analysis of this structure; such as the identification of local and global characteristics [22]. Community detection is one of these characteristics which appears when relationship between structure and function in networks has an important role. To do so; individuals or objects are represented by nodes and interactions among them are represented with edges, but in real world objects often have diverse roles and belong to multiple communities. In community detection, these objects should belong into multiple groups because they have multiple roles in the network, which are known as overlapping nodes. The aim of overlapping community detection is to discover such overlapping nodes and communities. Since this problematic often play important roles in network systems, many algorithms have been developed to solve it which can be roughly divided into two categories: heuristic methods (e.g. GN

[21]) and optimization based methods. In addition, nature has always been an inspiration for researchers; new nature-inspired algorithms have been developed to solve hard problems in optimization. In this paper, we use one of a nature-inspired algorithms bat algorithm to solve the problem of overlapping communities. The algorithm starts by the decoding step where individuals are represented with links of the network and optimized using partition density. After running the algorithm and having the best solution, the decoding step constructs communities and if two communities share one node, then they overlap. The algorithm is tested on real networks, and the obtained results show that bat algorithm have a higher density compared to other link clustering methods.

The paper is organized as follow: the section 2 analyses the state-of-art in overlapping community detection, Section 3 defines the proposed algorithm and section 4 explains how to apply the algorithm on the problem. In section 5 we define the objective function and in section 6 we show the experimental results. To conclude we introduce section 7.

2 Related work

To uncover the overlapping structure of networks, many algorithms have been proposed; almost two classes can be cited: node-based algorithms and link-based algorithms. In the first class, nodes of network are collected immediately; using their structure of information. Many firmly algorithms originate from this class, such algorithms that utilize the local expansion by optimizing a local benefit function: iLCD[15] (Intrinsic Longitudinal Community Detection) ,IS [5], LFM [4], MONC [6], CIS [8], OSLOM [9]. To detect overlapping structure with label propagation, algorithms enable for each node several labels, such as COPRA [12] and SLPA [7]. Some fuzzy community detection algorithms calculate for all communities the chance that they contains each node, such as the algorithm proposed by Gregory [3], spectral clustering framework based algorithm [10], etc…

The number of communities in fuzzy community detection algorithms should

be specified in anticipation [3], as CPM and SCP do. COPRA and SLPA can determine the number of community automatically.

Furthermore, the clique percolation can also be used for overlapping community detection. This method is based on rolling k-clique over the network through other cliques with k-1 common nodes. Since one node can participate in more than one community, overlap naturally occurs, like CPM [2] and EAGLE [14].

To detect overlapping communities, a lot of algorithms have been adopted

As well as genetic algorithms. The algorithm GA-Net+ [19] proposed by Pizzuti is based on nodes representation which detect overlapping communities without the need to know in advance the exact number of groups. Recently Ahn, Bagrow and Lehmann (ABL) [20] proposed the link-based method which is applied by Ye et al. for huge networks [13]. ABL converts the link communities into nodes communities based on the incident relationship between edges and nodes, after discovering the link community partition. According to GaoCD[1], this category is more powerful than node-based algorithm. For this reason we choose to utilize it in our approach.

3 BAT ALGORITHM

Bat algorithm is one of bio-inspired algorithms founded on swarm intelligence. The algorithm was developed in 2010 by Xin-She Yang [11]. Bats follow echolocation of bats by using sonar echoes to detect and avoid obstacles. It is generally known that sound pulses are transformed into frequency which reflects from obstacle. Yang [11] used three generalized rules for bat algorithms:

1) All bats use echolocation to sense distance, and they also guess the difference between food/prey and background barriers in some magical way.

2) Bats fly randomly with velocity $v_i$at position $x_i$with a fixed frequency f_{min}, varying wavelength and loudness $A_0$to search for prey. They can automatically adjust the wavelength (or frequency) of their emitted pulses and adjust the rate of pulse emission r depending on the proximity of their target.

3) Although the loudness can vary in many ways, it is assumed that the loudness varies from a large (positive) A0to a minimum constant value Amin. The body of algorithm is as follow:

Bat Algorithm

Objective function f (x) , x =(x1,...,xd)T
Initialize the bat population x_i (i = 1, 2,..., n) and v_i
Define pulse frequency fiat x_i
Initialize pulse rates $r_i$and the loudness A_i
while(t < Max number of iterations)
Generate new solutions by adjusting frequency,
and update velocities and locations/solutions
if (rand >r_i)
Select a solution among the best solutions
Generate a local solution around the selected best solution
end if
Generate a new solution by flying randomly
if (rand < Aiand f(x_i) < f(x_*))
Accept the new solutions
Increase $r_i$and reduce A_i
end if
Rank the bats and find the current best x_*
end while
Post process results and visualization

Bat Algorithm is formulated for continuous constrained optimization problems. [11] From the formulation of the Bat Algorithm and its implementation and comparison, we can see that it is a very promising algorithm. It is potentially more powerful than particle swarm optimization and genetic algorithms as well as Harmony Search. The primary reason is that BA uses a good combination of major advantages of these algorithms in some way. Moreover, PSO and harmony search are the special cases of the

Bat Algorithm under appropriate simplifications. In addition, the fine adjustment of the parameters α and γ can affect the convergence rate of the bat algorithm. Though the implementation is more complicate than many other metaheuristic algorithms; however, it utilizes a balanced combination of the advantages of existing successful algorithms with innovative feature based on the echolocation behavior of bats.

4 How to apply BAT ALGORITHM on community detection problem?

First we start by generating the initial population and then we apply bat algorithm to optimize it.

4.1 Generation of initial population:

The encoding step is based on edges of the network, where we consider **i** as the identifier of edges $i \in [0, m-1]$ and m as the length of bat. An individual l is represented as $l = \{l0, l1, \ldots, li, \ldots, lm-1\}$.

Where each lirandomly takes one of the adjacent edges of edge i. Adjacent means that a pair of links meets at a common vertex.

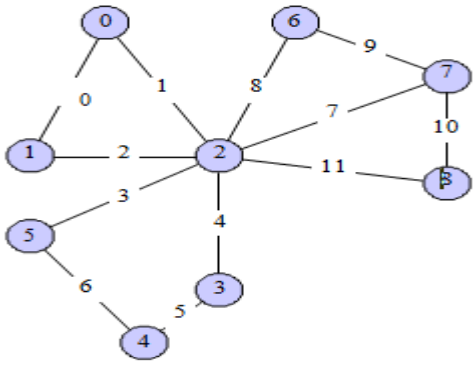

Fig. 1. Sample network.

Table 1 represents generation of bat according to the figure 1.

Table 1. Example of bat encoding.

Edge i	0	1	2	3	4	5	6	7	8	9	10	11
Adjacent edge to edge i	1	2	0	6	3	4	3	9	7	10	11	8

According to the definition of bat, two edges are adjacent if they share a common node. As a result we put it in the same partition. The process is repeated for all edge. Finally, we get a set of connected components which we call communities. Communities overlap if they share one or several nodes. To construct a community partition, the genotype is transformed shown in table 2:

Table 2. Community partition.

0	1	2	3	4	5	6	7	8	9	10	11
1	2	0	6	3	4	3	9	7	10	11	8

Then we get three communities:
C1={0,1,2} C2={3,4,5,6} C3={7,8,9,10,11}
The three communities share a common node 2.

5 Objective function:

The objective function is sometimes called the cost function. The space formed by this function is called the solution space. The value of objective function represents the importance of individuals, and guides the random search. Because we use links of the network to evaluate communities, we adopt partition density as objective function. Ahn et al. [20] proposed the partition density D, defined as follow: For a network with M links, we suppose P={P1,. . . ,Pc} as a partition of the network links into C subsets. mc= |Pc| is the number of links in subset c.
nc=| $\bigcup e_{ij} \in P_c$ {i, j}| represents the number of nodes incident to links in subset c. Dc refers to the link density of subset c, which is defined as follows:

$$D_c = \frac{m_c - (n_c - 1)}{\frac{n_c(n_c-1)}{2} - (n_c - 1)} \quad (1)$$

The partition density D is the average of Dcover all communities, weighted by the fraction of links present in each community. It is defined as follows:

$$D = \sum_c \frac{m_c}{M} D_c = \frac{2}{M} D_c = \frac{m_c-(n_c-1)}{\frac{n_c(n_c-1)}{2}-(n_c-1)} \quad (2)$$

6 Experiments

The benchmark used to show the efficacy of the method is the six much-usual real networks [17, 18].
Karate: Social network of friendships between 34 members of a karate club ata US university in the 1970s.
lesmis:Coappearance network of characters in the novel Les Miserables.
adjnoun: Adjacency network of common adjectives and nouns in the novel David-Copperfield by Charles Dickens.
football: Network of American football games between Division IA colleges during regular season Fall 2000.
dolphins: An undirected social network of frequent associations between 62dolphins in a community living o Doubtful Sound, New Zealand.
polbooks: A network of books about US politics published around the time ofthe 2004 presidential election and sold by the online bookseller Amzon.com.
Edges between books represent frequent copurchasing of books by the same buyers.
First we test bat algorithm on the six real networks, after that we make comparison between the partition density obtained by our method and others(ABL, GA-NET+ and GaoCD).
In second time we show the partition obtained after executing the algorithmon karate network.
The experiments are carried out 2.50 GHZ and 4G Ram computer. We set the parameters of the algorithm as follow:
[11]At iteration t the bat has a location x_i^*which is associated with a velocity v_i^t and frequency f_{min} . We denote x_* as the best solution f_{min} =0 and f_{max} =100.
The following rules explain how to calculate these parameters:

$$f_i = f_{min} + (f_{max} - f_{min})\beta \quad \beta \in [0,1] \tag{3}$$

$$v_i^t = v_i^{t-1} + (v_i^t - x_*)f_i \tag{4}$$

$$x_i^t = x_i^{t-1} + v_i^t \tag{5}$$

To generate a new solution we use equation (6) where A_t represents the loudness and $\epsilon \in$|-1,1|.

$$x_{new}^t = x_{old} + \epsilon A_t \tag{6}$$

The loudness A_i and the rate r_i are varying as follow:

$$A_i^{t+1} = \alpha A_i^t \tag{7}$$

$$r_i^t = r_i^0[1 - \exp(\gamma t)] \quad \alpha = \gamma = 0,9 \quad and \quad r_i^0 = 0 \tag{8}$$

Define d: length of solution is set as the number of links in each network. After that we execute the algorithm instruction per instruction, and at each iteration we adjust the parameters to choose the best individuals according to the objective function. After running the algorithm on the six networks we get the results illustrated in the table 3:

Table 3. Comparison of four link-based methods, according to the density D for each real networks.

Network	Karate	polBooks	Dolphins	football	lesmis	adjnoun
N° nodes	34	105	62	115	77	112
N° edges	78	441	159	613	254	425
			DENSITY			
Bat algorithm	0.7725	0.3120	0.5276	0.3130	0.9481	0.2804
GaoCD	0.5167	0.3115	0.4183	0.5815	0.6317	0.1421
GA-Net	0.4624	0.1926	0.3308	0.1507	0.5881	0.2361
ABL	0.2848	0.2867	0.3203	0.5500	0.5821	0.1310

From this table we can see that Bat algorithm has a higher density comparing with GaoCD, GA-Net and ABL for the most networks. This means that the communities have a denser links. For example, with dolphins network bat algorithm has 0.5276 Density where GaoCD have 0.4183 and 0.3 for ABL and GA-Net.

Analyzing karate network

The network is data collected from the members of a university karate club by Wayne Zachary in 1977. The ZACHE matrix represents the presence or absence of ties among the members of the club. After applying the method on the network, we get the following partitions:

C1= {2, 1,22}
C2={1, 2, 3, 4, 8, 14, 20, 31}
C3={1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34}
C4={3, 1,8}
C5={33, 15, 34}
C6={30, 27, 34}

We can see that the network is portioned in several communities which overlap in some nodes. For example, the node 1 is appearing in three different C1, C2, and C3 which make these communities overlap. Another node 34 is shared by three other communities C3, C5 and C6. We can see that the community C4 is included in C2. And C5, C6 are included in the community C3 which makes a hierarchy in the network.

7 Conclusion

The paper introduces a metaheuristic Bat Algorithm which showed more efficiency than the genetic algorithm. In this method individuals are represented by links of the networks and optimized by the density like objective function. The echolocation behavior of bats and the setting of parameters make bat algorithm very efficient and quickly converge to the best solution. The comparison of the density of resulting communities with the results of other methods shows that our method produces denser communities.

References

1. Chuan Shi, YananCai, Di Fu, Yuxiao Dong, Bin Wu, A link clustering based overlapping community detection algorithm, Data Knowledge Engineering 87 (2013) 394â404
2. G. Palla, I. Derenyi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005) 814â818.
3. S.H. Zhang, R.S. Wang, X.S. Zhang, Identification of overlapping community structure in complex networks using fuzzy c-means clustering, Physica A 374 (2007) 483â490.
4. A. Lancichinetti, S. Fortunato, J. Kertesz, Detecting the overlapping and hierarchical community structure in complex networks, New Journal of Physics 11 (2009) 033015.
5. J. Baumes,M. Goldberg, M. Krishnamoorthy,M.Magdon-Ismail, N. Preston, Finding Communities by Clustering a Graph into Overlapping Subgraphs, IADIS, 2005, 97â104.
6. F. Havemann, M. Heinz, A. struck, J. Glaser, Identification of overlapping communities and their hierarchy by locally calculating community-changing resolution levels, Journal of Statistical Mechanics: Theory and Experiment 01 (2011) 01023.
7. J. Xie, K. Szymanski, X. Liu, SLPA: Uncovering Overlapping Communities in Social Networks via a Speaker and Listener Interaction Dynamic Process, ICDMW, 2011, 344â349.
8. S. Kelley, The existence and discovery of overlapping communities in large-scale networks, Ph.D. thesis Rensselaer Polytechnic Institute, Troy, NY, 2009
9. A. Lancichinetti, F. Radicchi, J.J. Ramasco, S. Fortunato, Finding statistically significant communities in networks, PLoS One 6 (4) (2011) 18961.
10. S. Zhang, R.S. Wang, X.S. Zhang, Uncovering fuzzy community structure in complex networks, Physical Review E 76 (2009) 046103.
11. X.S. Yang. A new metaheuristic bat-inspired algorithm. Nature Inspired Cooperative Strategies for Optimization (NICSO 2010),pages 65â74, 2010.
12. S. Gregory, Finding overlapping communities in networks by label propagation, New Journal of Physics 12 (2010) 10301.
13. Q. Ye, B. Wu, Z.X. Zhao, B. Wang, Detecting Link Communities in Massive Networks, ASONAM, 2011, 71â78.
14. H.W. Shen, X.Q. Cheng, K. Cai, M.B. Hu, Detect overlapping and hierarchical community structure in networks, Physica A 388 (8) (2009) 1706â1712.

15. JieruiXie and Boleslaw K. Szymanski, Xiaoming Liu, âSLPA: Uncovering Overlapping Communities in Social Networks via A Speaker-listener Interaction Dynamic Process Â»
16. Xin-She Yang, A New Metaheuristic Bat-Inspired Algorithm, Department of Engineering, University of Cambridge.
17. http://www-personal.umich.edu/ mejn/netdata/
18. http://graph-tool.skewed.de/static/doc/dev/collection.html
19. C. Pizzuti, Overlapping Community Detection in Complex Networks, GECCO, 2009, 859â866.
20. Y.Y. Ahn, J.P. Bagrow, S. Lehmann, Link communities reveal multi-scale complexity in networks, Nature 466 (2010) 761â764. 2011, 344â349.
21. M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences of the United States of America 99 (2002) 7821â7826.
22. David F. Nettleton, Data mining of social networks represented as graphs, computer science review 7(2013) I-34.

Instance Matching Tools for Linked Data: A Comparative Study

Abderrahmane Khiat[1], Moussa Benaissa[1]

[1] LITIO Laboratory, University of Oran 1 Ahmed Benbella,
BP 1524 El-Mnaouar Oran, Algeria

abderrahmane_khiat@yahoo.com, moussabenaissa@yahoo.fr

Abstract. The objective of Linked Data with the emergence of the Web of Data is to interlink semantically data together in order to be reused and processed automatically by the software agents. These data described by instances are heterogeneous and distributed. The Instance matching is a very necessary task in Linked Data; it aims to identify the instances that describe the same real-world objects. The enormous volume of data already available on the web and its continuity to increase, requires techniques and tools capable to identify the instances that describe the same real-world objects automatically. In this paper, we describe InsMT-W an improved version of our InsMT system which participated in OAEI 2014. This second version consists to use WordNet in order to enrich instances with terminological information, extracted from documents that are relevant of the two instances to align. This method is considered as another way to annotate instances. We attempt to send also this second version to OAEI 2015 in order to be evaluated by OAEI campaign and get the official results.

Keywords: Semantic Annotation, Semantic Enrichment, Instance Matching, Semantic Annotation, Linked Data, Web of Data, Semantic Interoperability, Semantic Web.

1 Introduction

The current Web, contains documents in various formats (PDF, Excel, HTML file) connected by hypertext links, also known as the Web of Documents. In our case, we call "document" if the content is unstructured and not exploitable, in opposite we call "data" if the content is structured and exploitable.

The inadequacy of the Web of Documents resides in the fact that the content of these documents is probably unstructured which means that it is not exploitable and untreatable automatically in applications, either by the machine or by expressive queries.

Faced with these problems, and especially for the re-use and sharing of content, the transition from the document to the data is very necessary. This involves the use of semantic web technologies in order to (a) publish structured data on the Web, (b)

make possible, the links between data from one data source to data within other data sources. These two points are very important to ensure semantic interoperability.

These data should be expressed using the RDF language (Resource Description Framework [see section 2.1]) to achieve the two major points that we have mentioned in order to enable the semantic interoperability, which led to the emergence of the Web of Data.

The data structure in this form can be easily interpreted by the computer and re-used in applications and easily linked with other data. If the data are easily linked the computer can work through relationships with other data and in this case the interoperability will be ensured. Other advantages of Linked Data among others are: improving the data quality, less human intervention and processing and short development cycles (quicker and save time).

With the effort of Linked Data Community to publish existing open license datasets as Linked Data on the Web and interlink things between different data sources, the Web of Linked Data has seen remarkable increase over the past years.

In terms of statistics, in 2007, over 500 million RDF triples published on the web with around 120,000 RDF links between data sources. In 2010, over the 28.5 billion triples, in 2011 over 31.6 billion triples and in 2013 over 50 billion triples. According to these statistics, the Linked Data seems to be increasing drastically [6].

Linked Data, by definition, links the instances of multiple sources. A common way to link these instances to others is to use the owl:sameAs property.

The enormous volume of data already available on the web and its continuity to increase, requires techniques and tools capable to identify the instances that describe the same real-world objects automatically.

Otherwise, the OAEI evaluation campaign distinguishes between matching systems that have participated in the category of ontology matching and those that have participated in the category of instance matching, in order to evaluate their performance.

We see that few systems[1] [10] have participated to test their performance at instance matching track of OAEI 2014, only our InsMT system and RiMOM-IM succeed to finish all sub-tracks of instance matching track of OAEI 2014.

In this paper we deal with two challenges namely:

1. The distributed and heterogeneous natures of data described with instances and
2. The huge volume of data available on the web and its continuous increasing [14].

Indeed, the Solution to this problem consists to provide techniques and tools capable to identify the instances that describe the same real-world objects automatically.

In this paper, we describe InsMT-W the second version of our InsMT system in order to resolve the problem of instance matching automatically.

[1] The declaration of OAEI 2014 evaluation campaign about instance matching systems "Again, given the high number of publications on data interlinking, it is surprising to have so few participants to the instance matching track, although this number has increased".

In the first version of our InsMT system [8], we annotate instances with both names and labels of concepts that contain these instances and with the names and labels of properties related to these concepts.

Contrariwise in InsMT-W (The second version of our InsMT system) we annotate instances with terminological information extracted from documents that are relevant of the instances to align using WordNet[2]. In this way we enrich instances with terminological information which is very useful especially when the instances lack of this information.

The official results[3] obtained by running the first version of our system in instance matching track[4] are good in terms of recall compared to other systems that participated in OAEI[5] 2014, however the results but are not good in terms of precision and F-measure. With InsMT-W we attempt to improve these results in terms of precision, recall and F-measure.

The rest of the paper is organized as follows. First, preliminaries on instance matching are presented in section 2, the related work on instance matching systems that participated in Instance Matching Track of OAEI 2014 is presented in Section 3. In the Section 4 we describe our system by giving a detailed account of our approach. The implementation is presented in Section 5. The Section 6 contains concluding remarks and sets directions for future work.

2 Preliminaries

In this section, we present the basic notions of Instance Matching.

The Linked Data consist to relate data with typed links across the Web using URIs, HTTP and RDF. The linked Data principles are defined by Tim Berners-Lee in 2007 [11]. These principles are as follow:

- Use URIs as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful RDF information.
- Include RDF statements that link to other URIs so that they can discover related things.

[2] http://wordnet.princeton.edu/
[3] http://islab.di.unimi.it/im_oaei_2014/index.html
[4] http://oaei.ontologymatching.org/2014/
[5] OAEI (Ontology Alignment Evaluation Initiative) organizes evaluation campaigns aiming at evaluating ontology matching technologies. http://oaei.ontologymatching.org/

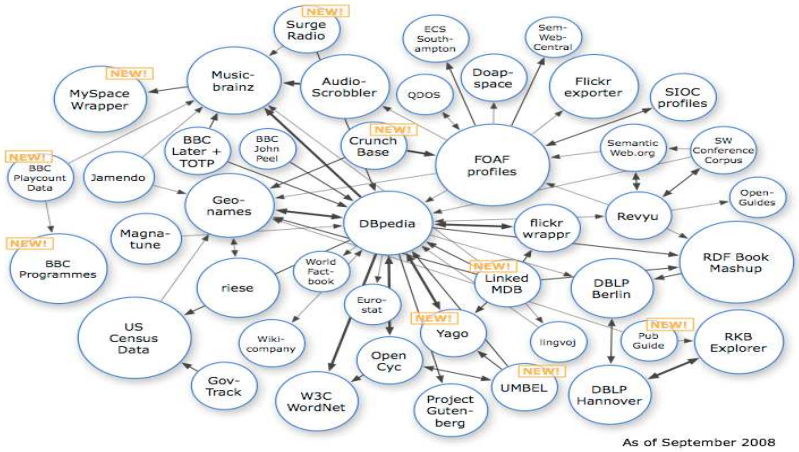

Fig. 1 Linked Data

Linked Data (fig. 1), by definition [12], links the instances of multiple sources. A common way to link the instances in these sources to others, is the use of the owl:sameAs property. Instance matching is required to interlink these data.

We give below the main notion related to linked data.

2.1 RDF Language

These data should be expressed using the RDF language (Resource Description Framework) to achieve the two major points that we have mentioned above in order to enable the semantic interoperability.

RDF language is a graph model to formally describe Web resources and metadata, in order to allow automatic processing of such descriptions [13][1][2]. An RDF file thus formed is a labeled directed multi-graph. Each triplet corresponds to a directed arc whose label is the predicate, the source node is the subject and the target node is the object.

We give below an example that shows how to link data from DBpedia with other data sources using the "owl:sameAs" property.

<http://dbpedia.org/resource/Berlin> owl:sameAs <http://sws.geonames.org/2950159>

2.2 Instance Matching

The *Instance Matching* (fig.2) is a process that starts from collections of data as input and produces a set of mappings (simple or complex) between entities of the collections as output [5].

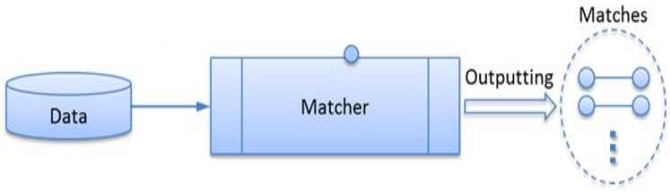

Fig. 2. Instance Matching.

2.3 Entity Resolution Notion [5]

Let D_1 and D_2 be represent two datasets, each one contains a set of data individuals T_i which are structured according to a schema O_i. Each individual $I_{ij} \in T_i$ describes some entity w_j.

Two individuals are said to be equivalent $I_j \equiv I_k$ if they describe the same entity $w_j = w_k$ according to a chosen identity criterion. The goal of the entity resolution task is to discover all pairs of individuals $\{(I_{1i}, I_{2j}) \mid I_{1i} \in T_1, I_{2j} \in T_2\}$ such that $w_{1i} = w_{2j}$.

In the context of linked data, datasets D_i are represented by RDF graphs. Individuals $I_i \in T_i$ are identified by URIs and described using the classification schema and properties defined in the corresponding ontology O_i.

3 Related Work

We present and discuss in this section the major works relevant to instance matching that participated at OAEI 2014 evaluation campaign.

Only two systems succeed to finish all sub-tracks of instance matching track of OAEI 2014, namely RiMOM-IM and our InsMT system.

We cite in exhaustive way only the instance matching systems that have participated in OAEI 2014 evaluation campaign and which are the object of comparison with our system InsMT-W.

1) LogMap [7]: The LogMap family participated with four different versions namely LogMap, LogMap-Bio, LogMap-C and LogMapLite in OAEI 2014. Only two versions (LogMap and LogMap-C) of them have participated at instance matching track. The principle of LogMap-family system is to create an initial alignment using lexical and structural indexing that operates on the ontologies to align. Then, it generates the final alignment by alternating iteratively the repair and discovery of new semantic correspondences using DL-reasoner. It incorporates both repair capabilities and discovery of semantic correspondences using reasoning-based approach.

The LogMap and LogMap-C systems finish only the first sub-track of instance matching of OAEI 2014 which is Identity Recognition.

2) RiMOM-IM [9] [3] [4]: is an acronym of **Ri**sk **M**inimization based **O**ntology **M**apping **I**nstance **M**atching. The principle of RiMOM-IM is to construct a document from the dataset by extracting the instances information. Then, it uses cosine-similarity to compare documents. The version of RiMOM-IM system that

participated in OAEI 2014 for instance matching is developed based on ontology matching system RiMOM with some changes in objective. The objective of RiMoM-IM is to solve the challenges in large-scale instance matching by proposing a novel blocking method.

3) InsMT [8]: is an acronym of **Ins**tance **M**atching at **T**erminological level. InsMT has participated for the first time in OAEI 2014. The principle of InsMT is to use String-based algorithms in order to calculate similarities between instances after the annotation step. The similarities calculated by each matcher are aggregated using the average aggregation strategy after a local filtering. Finally InsMT system operates a global filtering in order to identify the alignment. The InsMT system shows good results in terms of recall on different sub-tracks of instance matching of OAEI 2014.

The InsMT system finishes all sub-tracks of instance matching of OAEI 2014 which is Identity Recognition and Similarity Recognition.

4) InsMTL [8]: is an acronym of **Ins**tance **M**atching at **T**erminological and **L**inguistic level. InsMTL is our other system which has participated for the first time in OAEI 2014. The principle of InsMTL is to combine different terminological matchers with linguistic matcher (WordNet) by giving the priority to linguistic matcher otherwise the average aggregation is applied. Finally InsMTL system operates a filtering in order to identify the alignment. The InsMTL system shows also good results in terms of recall on different sub-tracks of instance matching of OAEI 2014. The InsMTL system finishes only the first sub-track of instance matching of OAEI 2014 which is Identity Recognition.

5) Other Approaches:

There are several other instance matching approaches like **HMatch** [18], **FBEM** [17], **SILK [16]** and the works proposed in [15] which are not covered by this paper due to minor importance for our approach. These instance matching approaches have not participated in instance matching track of OAEI 2014 and they are not covered by this paper due to minor importance for our approach. With respect to these approaches, we did not take them in consideration because we do not have their official results for the experimental protocol of OAEI in 2014.

As we have mentioned before, the declaration of OAEI evaluation campaign about instance matching systems with the high number of publications about interlinking approaches is surprising. Only RiMoM-IM and our InsMT system succeed to finish all sub-tracks (Identity Recognition and Similarity Recognition tasks) of instance matching track of OAEI 2014.

* In order to improving our InsMT system [8] we have used another way of annotation in our approach. We have annotated the instances with terminological information extracted from documents that are relevant of the instances to align using WordNet and then using different String-based methods to calculate similarities between these annotated instances. The advantage of our approach relative to others is that we enrich instances with terminological information obtained from the data that

are relevant to instances to be aligned. Our approach is very useful especially when the instances lack of this information.

4 Our Instance Matching Approach

The approach proposed in this paper is situated in **Terminological and Linguistic based methods** to resolve the problem of instance matching when instances to align lack form terminological information.

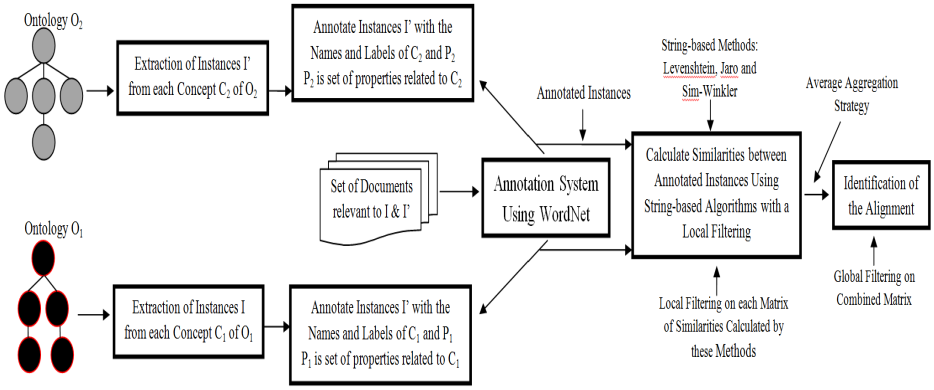

Fig. 3. Our Instance-based Ontology Matching.

Our approach aims first to annotate instances with the names extracted from the documents that are relevant of the instances to align using WordNet.

Then we use different string-based matchers in order to calculate similarities between these annotated instances with a local filtering.

Finally, we select the similar instances using a threshold S. if the similarity value is greater than the threshold, then the two instances are considered similar.

We summarize the process of our approach in Fig.3 to provide a general idea of the proposed solution. It consists in the following successive phases:

4.1 Phase 1: Annotation of Instances

In this phase, our system annotates the instances with the names extracted from the documents that are relevant to the instances to align using WordNet i.e. extraction of names from documents that are similar to instances to align using WordNet. The purpose of this annotation is to enrich the instances with terminological information. This step is very important especially when instances do not contain terminological information.

4.2 Phase 2: Calculation of Similarities

In this phase, our system calculates the similarities between instances, annotated in previous phase, using various string-based matching algorithms. The similarities calculated by each string matching algorithm are represented in matrix.

4.3 Phase 3: Local Filtering

In this phase, our system applies a local filtering on each matrix i.e. we choose for each string-based matching algorithm a threshold to realize a filtering. We set the similarities which are less than the threshold to 0. Our intuition behind this local filtering is that the similarities which are less than the threshold can influence the strategy of the average aggregation.

4.4 Phase 4: Identification of Alignment

In this phase, our system combines the similarities of different matrices (after we have applied a local filtering) using the average aggregation method and the result of the aggregation is represented in a matrix.

4.5 Phase 5: Global Filtering and Identification of Alignment

In this final phase, our system applies a second filtering on the combined matrix (result of the previous step) in order to select the correspondences found using the maximum strategy with a threshold.

5 Conception and Implementation

To implement our approach presented in the previous section, we propose the following architecture as shown in Fig. 4.

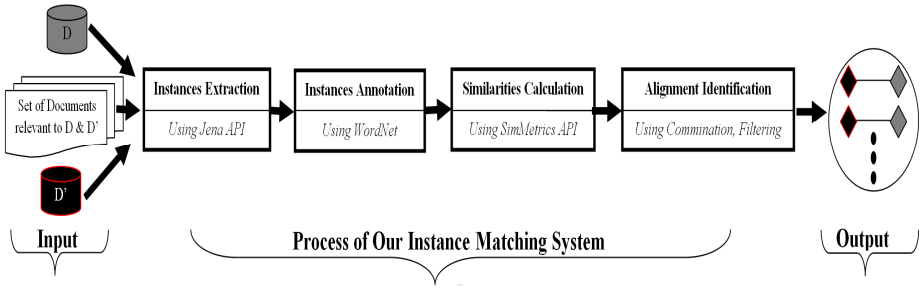

Fig. 4. Conception and Implementation of our Platform

The instances extraction phase was performed using Jena Plug-in of the Protégé platform. The instances annotation phase was achieved with WordNet. The tasks of calculating similarities and alignment extraction were performed respectively using

the different string-based methods namely: Levenshtein-distance, Jaro, and Slim-Winkler with a local filtering, strategy of average aggregation and global filtering for selection of instances that describe the same real-object.

6 Conclusion

In this article, we have introduced a new instance matching approach in order to identify the instances that describe the same real-world objects automatically. Our approach is useful especially when the instances lack of terminological information.

This is the second version of our InsMT system. The important part of the second version of InsMT system is the annotation of instances with terminological information extracted from the documents that are relevant of the instances to align using WordNet. Also it combines the various string-based matching algorithms in order to calculate similarities with average aggregation method. Finally we have applied a filtering on the combined matrix for the selection of instances that describe the same real-world objects.

The purpose of this annotation is to enrich the instances with terminological information.

The purpose of this new approach is that the first version of our InsMT system provides good results only in terms of recall but not in terms of F-measure. We attempt to participate with this new version of our system at OAEI 2015 in order to get better results than the first one.

References

1. J. Euzenat and P. Shvaiko "Ontology Matching", "Ontology Matching, Second Edition", Springer-Verlag, Heidelberg, pp. 1-511, 2013.
2. M. Ehrig "Ontology Alignment: Bridging the Semantic Gap", Semantic Web And Beyond Computing for Human Experience 4, Springer, pp. 1-250, 2007.
3. Z. Wang, X. Zhang, L. Hou,Y. Zhao, J. Li, Y. Qi and J. Tang "RiMOM results for OAEI 2010", In The Proceedings of the 4th International Workshop on Ontology Matching co-located with the 9th International Semantic Web Conference (ISWC 2010), pp. 195-202. CEUR-WS.org, Vol. 689, Shanghai, China, 2010.
4. J. Li, J. Tang, Y. Li and Q. Luo "RiMoM: a Dynamic Multistrategy Ontology Alignment Framework", Journal IEEE Transactions on Knowledge and Data Engineering, vol. 21, No. 8, pp. 1218-1232, 2009.
5. A. Ferrara, A. Nikolov, J. Noessner and F. Scharffe "Evaluation of instance matching tools: The experience of OAEI", Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 21 pp. 49-60, 2013.
6. C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), 2009.
7. E. Jiménez-Ruiz, B. C. Grau, W Xia, A. Solimando, X. Chen, V. Cross, Y. Gong, S. Zhang and A. Chennai-Thiagarajan "LogMap family results for OAEI 2014". In Proceedings of the 9th International Workshop on Ontology Matching co-located with the 13th International Semantic Web Conference (ISWC 2014), October 20, pp. 126-134. CEUR-WS.org, Trentino, Italy, 2014.

8. A. Khiat, M. Benaissa, "InsMT / InsMTL results for OAEI 2014 instance matching". In Proceedings of the 9th International Workshop on Ontology Matching co-located with the 13th International Semantic Web Conference (ISWC 2014), October 20, pp. 120-125. CEUR-WS.org, Trentino, Italy, 2014.
9. C. Shao, L. Hu and J. Li, "RiMOM-IM results for OAEI 2014". In Proceedings of the 9th International Workshop on Ontology Matching co-located with the 13th International Semantic Web Conference (ISWC 2014), October 20, pp. 149-154. CEUR-WS.org, Trentino, Italy, 2014.
10. Z. Dragisic, K. Eckert, J. Euzenat, D. Faria, A. Ferrara, R. Granada, V. Ivanova, E. Jiménez-Ruiz, A. O. Kempf, P. Lambrix, S. Montanelli, H. Paulheim, D. Ritze, P. Shvaiko, A. Solimando, C. Trojahn, O. Zamazal, B. C. Grau, "Results of the Ontology Alignment Evaluation Initiative 2014". In Proceedings of the 9th International Workshop on Ontology Matching co-located with the 13th International Semantic Web Conference, pp. 61-104. CEUR-WS.org, Trentino, Italy, 2014.
11. Tim Berners-Lee. Linked Data - Design Issues, 2006. http://www.w3.org/DesignIssues/LinkedData.html. 7, 26, 82.
12. R. Parundekar, C.A. Knoblock, J.L. Ambite, "Linking and building ontologies of linked data". In: Proceedings of the 9th International Semantic Web Conference (ISWC 2010). Shanghai, China, 2010.
13. G. Klyne and J. J. Carroll. "Resource Description Framework (RDF): Concepts and Abstract Syntax" - W3C Recommendation, http://www.w3.org/TR/rdf-concepts/, 2004.
14. P. Shvaiko and J. Euzenat. Ten challenges for ontology matching. In R. Meersman and Z. Tari, editors, On the Move to Meaningful Internet Systems: OTM 2008, volume 5332 of Lecture Notes in Computer Science, pp. 1164–1182. 2008.
15. D. Engmann and S. Maßmann. Instance matching with COMA++. In Proceedings of Datenbanksysteme in Business, Technologie and Web(BTW 07), pages 28–37, 2007.
16. J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data, In The Proceedings of 8th International Semantic Web Conference (ISWC 2009), A. Bernstein, D. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta, and K. Thirunarayan, editors, The Semantic Web - ISWC 2009, volume 5823 of Lecture Notes in Computer Science, pp. 650–665. Springer Berlin / Heidelberg, 2009.
17. H. Stoermer and N. Rassadko. Results of okkam feature based entity matching algorithm for instance matching contest of oaei 2009, 2009.
18. S. Castano, A. Ferrara, S. Montanelli, and D. Lorusso. Instance matching for ontology population. In Proceedings of the 16th Italian Symposium on Advanced Database Systems, pages 121–132, 2008.

A Modified Bat Algorithm for Mining Association Rule

Kamel Eddine Heraguemi[1], Nadjet Kamel[1] and Habiba Drias[2]

[1]Univ-Setif, Fac-Sciences, Depart.Computer Science,
Setif, Algeria,
[2]LRIA,USTHB
Algiers, Algeria

Abstract. In the recent years, association rules are being, inside of the data mining techniques, one of the most tools used to find relationships between the different attributes of the datasets. With the fast growth of stored data in our world, traditional algorithms become a big consumer of time and memory. Nowadays, researchers deal with ARM issues as an optimization problem. In this paper, we propose a modified bat algorithm for mining association rules called BAT-ARM which is based on bat inspired algorithm. The advantage of the bat algorithm is the combination of population-based algorithm and the local search, however, it is more powerful in local search. Our proposed algorithm is tested on several generic datasets with different number of transactions and items. The results are compared to FP-Growth algorithm results on the same datasets. BAT-ARM algorithm performs better than the FP-Growth algorithm in term of computation speed and memory usage.

Keywords: Association rules mining, ARM, Bat algorithm, optimization, FP-Growth algorithm, Support, Confidence

1 Introduction

Knowledge discovery in databases has attracted the attention of researchers because of the fast increasing of stored data. Great number of studies were dedicated to association rule mining [1]. ARM allows the generation of clear and practical rules in large databases. These rules show the correlations between items and attributes in the databases. Association rules are suited and useful for number of applications such as marketing, medical diagnostic and telecommunication. Association Rule generation was generally performed in two steps: the first step aims to get the frequent item-sets or frequent pattern in the database. The goal of the second step is to generate the association rules using the frequent pattern extracted in the first step . Mining the frequent patterns is probably the most important concept in data mining. It is proven to be an expensive process in term of time and space because the huge number of generated item-set .

Since 1993, when Agrawal, Imieliński and Swami. [1] Introduce the definition of association rules and frequent pattern; several exact algorithms were proposed to solve this problem. Each algorithm has its method to generate the frequent pattern and associations. Apriori [2] is the most popular algorithm. It generates the candidate item-set with a pruning mechanism. However, its major drawback resides in the important number of scans of the whole database, which slows down the response time.

In [3], Han, Jiawei and Pei, Jian and Yiwen propose a new exact algorithm called FP-growth. It mines frequent pattern and generates the association rule without candidate generation. It needs just two database scans. Eclat algorithm proposed in[4] to use the vertical representation of database to reduce the counting of the support. The main drawbacks of these conventional algorithms are: firstly, they are tooresources consuming especially with a huge number of transactions in the datasets when there is not enough of physical memory. Secondly, they need too much execution time when it comes to sizable databases. Motivated by the success of bat algorithm in various fields of data-mining, such as classification [5], clustering [6], combinatory optimization and Image processing [7], we present in this paper a new modified bat algorithm for association rule mining called BAT-ARM algorithm. Our approach was tested on several common datasets and compared to FP-growt algorithm results.

The remainder of this paper is organized as follows: In the next section we present a general background about association rule and we recall the original of the bat algorithm. In section 3 we present a brief review of the literature on ARM exaction using evolutionary algorithms. In section 4, we describe and discuss our proposed algorithm. In section 5 we present our experimental results. In the last section we conclude this paper and we present our future work.

2 Preliminaries

In this section we provide a brief description on association rules, including some main definitions.

2.1 Definitions

Formally, the association rule problem is defined as follow: Let $I = \{i_1, i_2,...,i_n\}$ be a set of literals called items, Let D be a set of database transaction where each transaction T contains a set of items. The problem of mining association rules aims to generate all rules having support and confidence greater than the user specified minimum support (called *minsup)* and minimum confidence (called *minconf)* respectively.

Support :***supp(X)****, it is the proportion of transactions of D that contains X, the support get its value in [0,1]*.

Confidence : **conf(X→Y)**, *its the proportion of the transaction that covering X and Y,* $0 \leq conf(X \rightarrow Y) \leq 1$.

Association rules is a rule of the form $A \longrightarrow X$ expressing the fact of if/then statement, if A then X. An association rule has two parts: an antecedent (if) and a consequent (then).

Almost association rule mining algorithms use the statistical analysis to determine usefulness and correctness of the rules. Generally all the exact methods use the support and the confidence. Meta-heuristic approaches use different measures, Support, Confidence,Gain and lift, and optimize them using the fitness function to evaluate the generated rule.

2.2 Original Bat algorithm

Yang [8] proposed a new and interesting meta-heuristic optimization technique called Bat Algorithm. This technique was proposed to behave as a band of bats tracking prey/foods using their echolocation. To model this algorithm, few rules are defined and presented in [8]. At the beginning, the bat population is initialized with **Erreur !** for each bat b_i at the time *t*. Let *T* be the number of iterations. As mentioned in [8], motion of virtual bats is done by updating their frequency, velocity and position as follow

$$f_i = f_{min} + (f_{min} - f_{max})\beta \quad (1)$$

$$v_i^t = v_i^{t-1} + [v_i^{t-1} - x^*]f_i \quad (2)$$

$$x_i^t = x_i^{t-1} + v^t_i \quad (3)$$

Where $\beta \in [0,1]$ is a random generated number and x^* is the current best solution which is located after comparing all the solutions among all the bats. For the local search, Yang [8] uses a random walk to generate a new solution for each bat b_i. First, a solution is selected among the current best solutions, then the random walk is applied on the bats that have their rates smaller than the random rate *rate* as follows:

$$x_{new} = x_{old} + \varepsilon A^t \quad (4)$$

where $\varepsilon \in [-1,1]$ is a random number and A^t is the average loudness of all the bats at the time *t*. At each iteration of the algorithm, the loudness A_i is reduced and the rate r_i is increased. as follow:

$$A^{ti} = \alpha A_i^{t-1} \quad (5)$$

$$r_i^{t+1} = r_i^0[1-exp(-\gamma^t)] \quad (6)$$

where α and γ are constants. At the initialization step of the algorithm, each bat has a different random loudness A_0 which is in [1,2] and random rate r_0 which is in [0,1] as mentioned in [8].

3 Related work

In the literature, we can find a large set of techniques based on evolutionary algorithms proposed to extract association rule. Such type of algorithms are known to generate solutions using techniques inspired from natural behaviors such as Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and more other bio-inspired algorithms. All of them use initially a population of candidate solutions and evolve them to obtain the best solution to the problem.

Association rule mining can be considered as an optimization problem. So evolutionary algorithms are used to generate the best rules in dataset by selecting the rules that maximize the fitness function. In the literature, different works propose genetic algorithms with different fitness and genetic operation. Quantminer [9] is a

genetic algorithm for mining quantitative association rules. In this algorithm individuals represent rules and the algorithm evolves to search for the best solution. In[10], the authors propose to optimize association rule mining using new fitness function. This fitness function divides the rules into discrete and continuous ones. In [11], Yan and Zhang. propose a genetic algorithm for identifying association rules without specifying minimum support called ARMGA. the main inconvenience with this algorithm is the invalid chromosomes generated and the production of many rules.

On the other hand, AR mining can be seen as a multi-objective optimization problem, in which different measures are used in the fitness function to evaluate the rules. In this area, several works are proposed. In [12], a multi-objective genetic algorithm approach to mine association rules for numerical data was proposed, where confidence, interestingness and comprehensibility are used to define the fitness function. In[13], Qodmanan also propose a multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. A review on multi-objective rule mining is presented in [14].

In addition to genetic algorithm, there are many other bio-inspired approaches which are proposed to extract association rules. In [15], the authors propose a particle swarm algorithm to detect association rules. The algorithm is defined through two main parts: preprocessing and mining. The preprocessing part calculates the fitness of the particle swarm. In the mining part, PSO is used to mine the rules. A great review on Application of Particle Swarm Optimization in Association Rule Mining was proposed in [16].The authors of [17] developed a new algorithm called BSO-ARM. This algorithm is inspired from bees behavior and based on BSO algorithm. The results show that BSO-ARM performs better than all genetic algorithms. As extension to their work, the authors present improvements to BSO-ARM in [18], where three strategies to determinate the search area of each bee are proposed (modulo, next, syntactic). These improvements give its fact on the quality of rules extracted by BSO-ARM, but in the same time the algorithm takes more CPU time. The same authors present another Hybrid approach called (HBSO-TS)in[19] for mining association rules based on Bees swarm algorithms and Tabu-search. The results show that HBSO-TS extract useful rules in reasonable time.

4 BAT-ARM Algorithm

In this section we define the solution encoding we use in our approach, the fitness function and how bats ranks from position to another (Bat motions).

4.1 Encoding

In the literature, several representation of the rule to mine the ARM using genetic algorithms or meta-heuristic algorithms can be found. In ARMGA algorithm [11], the rule is represented as a chromosome of k+1 length, where the first position is the cut point that separates the antecedent and the consequent of the rule. Positions 1 to k are the index of the items, Let the rule X→Y, $X= \{A_1,\ldots, Aj\}$and$Y= \{A_{j+1},\ldots,A_K\}$. Another representation is considered by [20] where all items are appeared in the chromosome.

So, if we have n item then we get 2n length chromosome. In this representation, we have four codes that can be assigned to each item: 00: if the referred item is in the antecedent of the rule, **01** or **10**: the referred item is not included in the rule, **11**: if the referred item is in the consequent of the rule. For our approach, we use the first representation where each solution X represents a rule and contains k item. Therefore, the solutionX contains k+1 positions. Assume $I=\{I_1,I_2,\ldots,I_n\}$ the set of all items in the database. The rule X→Y is encoded as follow:

j	I_1	....	I_j	I_{j+1}	....	I_n

Where the j separates between the antecedent and the consequent of the rule, if i^{th} item in I is in the rule then the position k contains i else the position contains 0 where $0<k<=n+1;$

4.2 Fitness function

As mentioned above, in association rule mining, the rule is accepted if its support and confidence satisfy user minsup and minconf. Each work on evolutionary algorithms for mining association rule uses its fitness function for the individuals. Our fitness function is described as follow:

$$F(R) = \begin{cases} \alpha \times conf(R) + \beta \times supp(R)/(\alpha+\beta) & if\ accepted\ rule \\ -1 & otherwise \end{cases}$$

where $R=X \Longrightarrow Y$, and let α and β be two empirical parameters.

4.3 Virtual Bat motion

In [8] the authors present a mathematical simulation of the natural bat movement, where the frequency f_i, the velocity v_i and the position (solutions) x_i for each virtual bat are described, and these values are generated based on three equations presented in 1, 2, and 3. In our approach we propose a new description for bat motion related to association rule mining. The same concepts are defined: frequency, velocity and position.

- **Frequency $\mathbf{f_i}$**: presents how much this bat is strong to change its position. In other words, it means how many items can be changed in the actual rule, where the maximum frequency f_{max} is the number of attributes in the dataset and the minimum frequency f_{min} is 0.
- **Velocity $\mathbf{v_i}$**: presents where the changes will be started.
- **Position $\mathbf{x_i}$**: it is the new generated rule based on new frequency, velocity and the loudness .

The new generated rule at the iteration t is given by:

$$f_i^t = 1+(f_{max})\beta \quad (7)$$

$$v_i^t = f_{max} - f_i^t - v_i^{t-1} \quad (8)$$

Algorithm 1 Generate new solution

Input: Rule x_{i-1}, Frequency f_i, velocity v_i, loudness Ai
Output: new rule x_i

```
While (v_i< Frequency f_i) do
If (rand > A_i) then
        Item at v_i ← Item at v_i + 1
Else
        Item at v_i ← Item at v_i - 1
End if
If (Item at v_i less or equal to 0 or greater than number of attributes) then
    Item at v_i ← 0
End if
If The new rule contains duplicated items then
    Keep one randomly and delete the others.
End if
Increment v_i
End while
```

The new position (rule) x_i is generated by applying **algorithm 4.3**, such that, if the loudness is less than a random value *rand* we increase the value of actual bit, else we decrease this value. If the value is out of the interval [1..N], the value is replaced by 0. The new generated rule may contains duplicated bits (items). In this case, we keep a random bit from the duplicated and delete the others. Here we never get an invalid rule. This operation will be repeated starting from the bit at current velocity to achieve the actual frequency. The modified bat algorithm pseudo code is presented in **Algorithm 2.**Erreur ! Source du renvoi introuvable.

Algorithm 2 BAT-ARM Algorithm

```
Objective function (fitness function)
Initialize the bat population x_i and v_i
Define pulse frequency f_i at x_i
Calculate the fitness of each initial position f_iat x_i
Initialize pulse rates r_i and the loudness A_i
While (t < Max number of iterations) do
    Generate new solutions by adjusting frequency f_i,
    and updating velocities and locations/solutions [equations 7, 8]
    Generate a new solution x_i using Algorithm 1 where inputs are f_i, v_i, A_i
    If (rand >r_i) then
        Generate a local solution around the selected best solution by changing just one
        item in the rule
    End if
    If (f(x_i) > f(x_i*)) then
        Accept the new solutions
        x_i* = x_i
        Increase r_i and reduce Ai equation 5, 6
    End if
```

Rank the bats to the best solution
End while
Post-process results and visualization the best detected rules

4.4 Complexity of BAT-ARM

Association rule mining is one of the most atractive problem in NP-class [21]. For the proposed algorithm at each iteration, each bat i of the n bats generates a new solution starting from rule that contains k bits (items), with the use of Algorithm 4.3. So the number of modified bits is the $frequency_i - velocity_i$. In the worst case the frequency will be equal to F_{max} and the velocity is 0, then the items will be changed. Here the complexity is $O(n \times Maxiterations \times k)$.

5 Experimentation and results

To evaluate our algorithm we tested it on a several synthetic common standard databases with different number of transactions and items. The programs are written in Java and run on intel core I5 machine with 4Go of memory running on linux ubuntu. Results are described in the following section.

In our experiment, we change the two essential parameters in our algorithm: the number of bats *n* and the number of iterations *t*, and we fixed α and β to 1. As mentioned in [8], α and γ in ***equation 5,6*** are α=γ=0.9.

Table 1 presents the results of running our proposed algorithm on synthetic common database of 1000 transactions and 40 items were created with the IBM dataset generator[2]. The initial positions are generated randomly, and the minimum support and the minimum confidence are fixed to 0.2 ad 0.5 respectively. The results show that our algorithm yields a good performance in term of time and memory usage with max time 19 seconds and 2.74 Megs of memory usage. The number of iterations is 200 and the number of bats is 50. On the other hand, FP-growt algorithm is more efficient in term of time with 0.73 second with higher memory usage.

Table 1: BAT-ARM results with common database (1000 transactions and 40 items)

Bats number	iteration	execution time (s)	best fitness	Memory usage (Megs)	FP-growt runtime(s)	Memory usage (Megs)
10	25	0.56	0.3795	1.31	0.73	16.97
	50	1.06	0.4039			
	100	2.03	0.4087			
	150	2,9	0.4087			
	200	3	0.4118			
25	25	1	0.3815	2.45	0.73	16.97

	50	2	0.4138			
	100	5	0.4137			
	150	7	0.4162			
	200	9	0.4087			
	25	2	0.4087			
	50	5	0.4087			
50	100	10	0.4087	2.74	0.73	16.97
	150	15	0.4118			
	200	19	0.4138			

Table 2 shows the results of running our tests on CHESS[22] database that contains 3,196 transactions and 75 items. The results show that our algorithm gives a good runtime with a good quality of rules generated and less memory usage. Our algorithm goes to maximize the fitness function till 0.9694 with support 0.92 and confidence 0.93 and maximum memory usage 48.63. For FP-growt algorithm the runtime grows up because of the number of transactions in the CHESS dataset.

Table 2: BAT-ARM results with Chess database (3,196 transactions and 75 items)

Bats number	iteration	execution time (s)	best fitness	Memory usage (Megs)	FP-growt runtime(s)	Memory usage (Megs)
	25	2	0			
	50	9	0.7034			
10	100	13	0.8044	42.51	523	104.55
	150	22	0.8209			
	200	28	0.8152			
	25	12	0			
	50	38	0.632			
25	100	72	0.6464	45.95	523	104.55
	150	60	0.8728			
	200	73	0.8152			
	25	34	0.5795			
	50	38	0.8837			
50	100	67	0.8565	48.63	523	104.55
	150	125	0.9044			
	200	141	0.9295			

Table 3 shows the results of our tests on mushroom[22] database that contains 8124 transactions and 119 items. Again, the results confirm the efficiency of our algorithm with the growth of the number of transactions and items in database in term of execution time and memory storage. The maximum usage of memory was 170 Mo when the parameters are 200 iterations and 50 bats. On the other hand, FP-growt algorithm goes till 291 Megs of memory and runtime of 1165 seconds.

Table 3: BAT-ARM results with Mushroom database(8124 transactions and 119 items)

Bats	iteration	execution	best	Memory	FP-growt	Memory usage

number		time (s)	fitness	usage (Megs)	runtime(s)	(Megs)
	25	6	0			
	50	10	0.6063			
10	100	26	0.6329	128.67	1165	291.1
	150	41	0.6927			
	200	68	0.6927			
	25	18	0			
	50	37	0.5622			
25	100	86	0.6329	156.69	1165	291.1
	150	115	0.6927			
	200	199	0.7070			
	25	16	0.6758			
	50	61	0.6464			
50	100	151	0.6927	170.29	1165	291.1
	150	253	0.7198			
	200	341	0.7376			

Based on the conducted experiments, we observe that our algorithm provides a great performance in term of CPU-time and memory usage. Thanks to the echolocation concept of the bat algorithm that can determine which part of the best rule have changed to get a better position (rule) for the actual bat.

6 Conclusion

In this paper, we presented a new application of bat algorithm for association rule mining. The proposed approach is inspired from bats behavior and based on echolocation concept to generate new positions (rules). Compared to FP-growt algorithm, our proposal proved its efficiency in term of time, memory usage and quality of generated rules with maximizing the fitness function. The main drawback in our algorithm is the lack of communication between bats that can generate redundancies in the positions while searching a new one. As future work, we plan to propose a cooperative bat algorithm for mining association rules.

References

1. R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in ACM SIGMOD Record, vol. 22, pp. 207–216, ACM, 1993.
2. R. Agrawal, R. Srikant, et al., "Fast algorithms for mining association rules," in Proc. 20th int. conf. very large data bases, VLDB, vol. 1215, pp. 487–499, 1994.
3. J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in ACM SIGMOD Record, vol. 29, pp. 1–12, ACM, 2000.
4. M. J. Zaki, "Scalable algorithms for association mining," Knowledge and Data Engineering, IEEE Transactions on, vol. 12, no. 3, pp. 372–390, 2000.
5. S. Mishra, K. Shaw, and D. Mishra, "A new meta-heuristic bat inspired classification approach for microarray data," Procedia Technology, vol. 4, pp. 802–806, 2012.

6. K. Khan, A. Sahai, and A. Campus, "A fuzzy c-means bi-sonar-based metaheuristic optomization algorithm," IJIMAI, vol. 1, no. 7, pp. 26–32, 2012.
7. X.-S. Yang and X. He, "Bat algorithm: literature review and applications," International Journal of Bio-Inspired Computation, vol. 5, no. 3, pp. 141–149, 2013.
8. X.-S. Yang, "A new metaheuristic bat-inspired algorithm," in Nature inspired cooperative strategies for optimization (NICSO 2010), pp. 65–74, Springer, 2010.
9. A. Salleb-Aouissi, C. Vrain, and C. Nortet, "Quantminer: A genetic algorithm for mining quantitative association rules." in IJCAI, vol. 7, 2007.
10. R. Haldulakar and J. Agrawal, "Optimization of association rule mining through genetic algorithm." International Journal on Computer Science & Engineering, vol. 3, no. 3, 2011.
11. X. Yan, C. Zhang, and S. Zhang, "Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support," Expert Systems with Applications, vol. 36, no. 2, pp. 3066–3076, 2009.
12. B. Minaei-Bidgoli, R. Barmaki, and M. Nasiri, "Mining numerical association rules via multi-objective genetic algorithms," Information Sciences, vol. 233, pp. 15–24, 2013.
[13]H. R. Qodmanan, M. Nasiri, and B. Minaei-Bidgoli, "Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence," Expert Systems with applications, vol. 38, no. 1, pp. 288–298, 2011.
14. S. Srinivasan and S. Ramakrishnan, "Evolutionary multi objective optimization for rule mining: a review," Artificial Intelligence Review, vol. 36, no. 3, pp. 205–248, 2011.
15. R. J. Kuo, C. M. Chao, and Y. Chiu, "Application of particle swarm optimization to association rule mining," Applied Soft Computing, vol. 11, no. 1, pp. 326–336, 2011.
16. S. Ankita, A. Shikha, A. Jitendra, and S. Sanjeev, "A review on application of particle swarm optimization in association rule mining," in Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), pp. 405–414, Springer, 2013.
17. Y. Djenouri, H. Drias, Z. Habbas, and H. Mosteghanemi, "Bees swarm optimization for web association rule mining," in Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on, vol. 3, pp. 142–146, IEEE, 2012.
18. Y. Djenouri, H. Drias, and Z. Habbas, "Bees swarm optimisation using multiple strategies for association rule mining," International Journal of Bio-Inspired Computation, vol. 6, no. 4, pp. 239–249, 2014.
19. Y. Djenouri, H. Drias, and A. Chemchem, "A hybrid bees swarm optimization and tabu search algorithm for association rule mining," in Nature and Biologically Inspired Computing (NaBIC), 2013 World Congress on, pp. 120–125, IEEE, 2013.
20. S. Dehuri, A. Jagadev, A. Ghosh, and R. Mall, "Multi-objective genetic algorithm for association rule mining using a homogeneous dedicated cluster of workstations.," American Journal of Applied Sciences, vol. 3, no. 11, 2006.
21. F. Angiulli, G. Ianni, and L. Palopoli, "On the complexity of mining association rules." in SEBD, pp. 177–184, 2001.
22.B. Goethls and M. J. Zaki, "Frequent itemset mining dataset repository," http://fimi.ua.ac.be/data/, 2003.