Topics of Interest

- **Information Technology**
  - Information Systems
  - IT Applications & Services
  - IT Platforms: Software & Hardware Technology
  - IT Strategies & Frameworks

- **Communication Technology**
  - Communication Devices
  - Communication Theory
  - Mobile Communications
  - Optical Communications
  - Satellite Communications
  - Signal / Image / Video Processing

- **Network Technology**
  - Computer & Communication Networks
  - Wireless Networks
  - Network Management
  - Network Security
  - NGN Technology
  - Security Management
This Page intentionally left blank.
A Sub-Optimal Policy for Connection Admission Control Mechanism in Cognitive Radio Sensor Networks

Elahe S. Hosseini
Department of Computer Engineering
University of Science and Technology
Tehran, Iran
elahe_hosseini@iust.ac.ir

Reza Berangi
Department of Computer Engineering
University of Science and Technology
Tehran, Iran
rberangi@iust.ac.ir

Received: October 24, 2016- Accepted: March 19, 2017

Abstract—Satisfying the quality of service (QoS) is a crucial issue in cognitive radio sensor networks (CRSNs) due to the highly variable nature of cognitive radio channels. Connection admission control (CAC) is a beneficial approach to manage the traffic to provide desired QoS. A CAC is proposed in this paper to optimize the packet loss ratio, jitter of packets and end to end delay in CRSNs. The proposed CAC decides based on the priority of data flows, network state and number of available channels. An estimation formula is proposed through a graph coloring approach to evaluate the required number of channels of network states. The proposed CAC is modeled by a semi Markov decision process (SMDP) and a sub-optimal policy is obtained by a value iteration method to achieve the maximum reward in network. Simulation results demonstrate that the proposed mechanism outperforms the recent proposed admission control mechanism in CRSNs.

Keywords- Cognitive radio sensor networks; admission control; QoS; semi Markov decision process (SMDP);

I. INTRODUCTION

Dynamic spectrum access (DSA) is one of the main solutions to use the spectrum in wireless networks efficiently. The cognitive radio (CR) is a precious technology to provide DSA in order to solve the spectrum scarcity problem. The primary users (PUs) are the licensed users which have higher priority to use channels. The CR-equipped users can use the unlicensed spectrum bands in the absence of PUs according to basic cognitive radio operations: spectrum sensing, spectrum decision and spectrum handoff [1]. A CR user senses the channels periodically (spectrum sensing), if a PU enters into its licensed channel, the CR user leave the channel immediately in order to minimize the interference on the transmission of PUs (spectrum handoff) and decide to select another free channel (spectrum decision) [1].

There are some applications such as industrial control and surveillance in wireless sensor networks which have some specific features such as delay sensitivity and burst traffic. With regard to these features and the requirements of wireless sensor networks, these networks can use the benefits of the CR technology in order to satisfy these requirements and to overcome the spectrum scarcity problem. The wireless sensor networks with CR-equipped nodes are entitled as cognitive radio sensor networks (CRSNs)
Admission control is a crucial mechanism for providing QoS when there are many requesting users to access the network with the limited resources simultaneously. The connection admission control (CAC) is a pro-active congestion control which estimates the network resources and then decides about data flows transmission.

There are some studies on CAC in cognitive radio networks (CRNs). The authors of [3] considered a joint admission control and channel allocation using a Markov decision process to support the delay sensitive communications of CR users. In [4], three admission control schemes are proposed using discrete-time Markov chain to minimize the forced termination probability of CR users. A joint admission control, eviction control and bandwidth management framework is proposed in [5] using semi Markov decision process. In [6], a CAC framework is proposed based on channel reservation for CR users and the buffer size of handoff operation in order to analyze the dropping and blocking probabilities. The authors of [7] considered joint admission control, scheduling and spectrum handoff in order to improve the performance of multimedia transmissions using a Markov model. These studies proposed some admission control schemes along with cognitive channel allocation, or scheduling, or spectrum handoff or bandwidth management that are related to admission control in the lower layers of the network. However, the connection admission control mechanisms in higher layer focus on the data flows and prefer to send fewer valuable data flows reliably rather than to send several data flows incompletely. This feature of connection admission control leads to improve the event reliability in CRSNs.

To the best of our knowledge, there is only one study on connection admission control in CRNs which is [3]. The proposed CAC in [3] is based on the correlations of data flows and the traffic characteristics of CRNs. In [3], the admission control mechanism decides based on the average capacity of CR channels and defined event reliability metric. The proposed admission control mechanism in [3] estimates the network resources on average and does not decide based on the considering of network state at each decision instance.

The contribution of this study is the propose of a CAC mechanism in CRNs based on the priority of data flows and the required resources of each data flow and also, the network state which composes of the number of active PUs, the ID and then the number of flowing CR sensors at each decision instance. This mechanism is formulated as a semi Markov decision process (SMDP) in order to reach an optimal decision making framework for each state during network lifetime. In the proposed mechanism, the number of required channels for each data flow is estimated by a graph coloring approach at each decision instance. According to this resource estimation, the network state and the optimal decision at each state are determined. The aim of this admission control is to send the maximum number of valuable data flows by considering the available network resources at each decision instance. On the other hand, when PU activity is high and the network resources are limited, sending a few valuable data flows is desirable in order to inform more valuable information of event toward the sink. The optimal decision policy of the proposed SMDP model is obtained through value iteration method. The simulation results represent the superiority of the proposed CAC mechanism over the last proposed admission control in CRSNs in the terms of packet loss probability, end to end delay and jitter.

The rest of this paper is organized as follows. Section 2 states the system model. The problem definition, formulation and solution are explained in Section 3. Simulation results are presented in Section 4, and finally, the paper concludes with some remarks in Section 5.

II. SYSTEM MODEL

This paper considers a cognitive radio sensor network with three types of nodes, CR sensor nodes, CR relay nodes and a sink node that are placed within a certain finite area to provide multiple views. The number of CR sensor users, CR relay nodes and PUs are considered as \( N_S \), \( N_R \) and \( N_{PU} \), respectively. With regard to the occurred event in the event area, some sensors request to send a data flow toward the sink. According to the physical conditions of the event and sensor nodes such as sensors' location, the distance of sensors to occurred event, and also their angle of view to the sensing area, the induced data flows of different sensor nodes have different importance. Therefore, the different weights are assumed for requesting data flows to send that are obtained by the proposed weighting scheme in [3]. It is assumed that these sensors generate constant bit rate (CBR) data flows [8]. The sink node has the knowledge about sensor nodes to decide on the admission of data flows. A CR node has two main operating modes: sensing mode and operating mode. First, a CR node senses the licensed spectrum to decide whether it is idle or occupied by a PU. Sensing time and sensing frequency are denoted by \( t_s \) and \( f_s \), respectively [9]. After sensing, the CR node enters in operating mode and sends data in a licensed spectrum channel if it is free of PU. The PUs activity is modeled as exponentially distributed inter-arrivals thus their arrival to their related channels is independent. The traffic of a PU can be modeled as a two-state arrival-departure process with arrival rate \( r_a \) and departure rate \( r_d \). A PU has two states: ON and OFF [10]. The ON state represents the period that PU operates on a channel, and CR node cannot use the channel. The OFF state represents the period that the PU does not operate on a channel, and CR nodes can use the channel. There are CH cognitive channels with the same bandwidth. For each channel, there is a PU (\( N_{PU} \neq CH \)) and all of the CR channels have similar PU activity. In each channel, A PU operates based on its arrival rate \( r_a \) and departure rate \( r_d \). When a PU starts to operate on its licensed channel, the operations of each active CR node on the licensed channel in the CRSN will be stopped. In other words, the activity of all CR nodes in the CRSN is affected by the PUs activity.
Table 1. Notation Table

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_S, N_{P_R}, N_{P_F}$</td>
<td>Number of CR sensor users, primary users (PUs), CR relay nodes</td>
</tr>
<tr>
<td>$CH$</td>
<td>Number of CR channels</td>
</tr>
<tr>
<td>$r_{a}, r_{d}$</td>
<td>Average arrival rate of each PU to the channel, average departure rate of PU from the channel</td>
</tr>
<tr>
<td>$w_i$</td>
<td>Weight of the $i^{th}$ sensor node</td>
</tr>
<tr>
<td>$r_i$</td>
<td>Rate of the $i^{th}$ sensor node</td>
</tr>
<tr>
<td>$n(t)$</td>
<td>Admission condition vector of the flows at decision epoch $t$</td>
</tr>
<tr>
<td>$a(t)$</td>
<td>Admission decision vector at decision epoch $t$</td>
</tr>
<tr>
<td>$q(t)$</td>
<td>Number of active PUs in the network at decision epoch $t$</td>
</tr>
<tr>
<td>$s(t)$</td>
<td>Network state at decision epoch $t$</td>
</tr>
<tr>
<td>$P_{u}$</td>
<td>Probability of using route $d$ related to sensor $i$</td>
</tr>
<tr>
<td>$K_i$</td>
<td>Number of possible routes between the sensor node $i$ and the sink node</td>
</tr>
<tr>
<td>$\Omega(n)$</td>
<td>Minimum number of required channels in each possible routes configuration</td>
</tr>
<tr>
<td>$\gamma(n)$</td>
<td>Optimal average required number of channels at state $s = (n, q)$</td>
</tr>
<tr>
<td>$P_s(a)$</td>
<td>Probability of transition from state $s$ to state $x$ by selecting the action $a$</td>
</tr>
<tr>
<td>$m_s(a)$</td>
<td>Decision variable of selection the action $a$ at the state $s$</td>
</tr>
<tr>
<td>$f_1$</td>
<td>Function of mapping the state space to the acceptable action space</td>
</tr>
<tr>
<td>$\tau_d(a)$</td>
<td>Average time after the action $a$ is selected in state $s$ until the next decision epoch (sojourn time)</td>
</tr>
<tr>
<td>$R(s,a)$</td>
<td>Earned reward at the state $s$ and selection of the action $a$</td>
</tr>
<tr>
<td>$\psi(CR)$</td>
<td>Worthless CR user who is transmitting data packets toward the sink node</td>
</tr>
</tbody>
</table>

III. PROBLEM DEFINITION AND FORMULATION

In a cognitive radio sensor network, several sensors are deployed in the event area to provide multiple observations of an event. When an event occurs, depending on the event place and sensing radius, some of the sensor nodes send data flows toward the sink node. Due to the constraints of the cognitive channels, sending all of these flows cannot be reasonable. Furthermore, it is needed to inform the sink node some information about the event. Therefore, a connection admission control is needed to provide the QoS of the cognitive radio sensor network.

The SMDP is a powerful tool in analyzing stochastic decision control processes satisfying Markov features with random decision epochs. The SMDP has a lot of potential applications in telecommunication, reliability control and maintenance [11]. In an SMDP, the system is in one of the states of a finite state set in each decision epoch. There is a finite action set for each state. The system state evolves in different the decision epochs according to a transition probability matrix which depends on the current system state and selected action from the action set. According to the selected action in each state transition, a cost/reward is obtained. The aim is to optimize the long-term average cost/reward [11].

With regard to SMDP properties, the considered problem and network assumptions, the appropriate theory to model the decision making process for this admission control is SMDP. It is necessary to identify SMDP components related to this problem that are introduced in the next subsections. The notations which are used in this model are listed in Table 1.

A. State Space

The system state represents some network information at the beginning of each decision epoch. Define row vector $n(t) = [n_1(t), n_2(t), ..., n_N(t)]$ where $n_i(t) \in \{0, 1\}$ denotes the admission condition of the induced data flow from sensor $i$ in the event area at the decision epoch. The $n_i(t) = 1$ represents the sensor node $i$ has been admitted to send and is sending data flow toward the sink node. Also, the $n_i(t) = 0$ represents the sensor node $i$ has not been admitted to send data. Define $q(t)$ as the number of active PUs in the network at the decision epoch $t$. The network state is given by $s(t) = (n(t), q(t))$ at the decision epoch $t$ and also, is given by $s = (n, q)$ in steady state. The average number of required channels for each network state is considered as function $\gamma(n)$. Thus, the number of used channels by admitted flows plus active PUs should be less than $CH$. Therefore, the state space $S$ can be defined in Equ. 1.

$$S = \left\{ s = (n, q); n_i \in \{0, 1\}, 0 \leq q \leq CH, \gamma(n) + q \leq CH \right\}$$

The details of the function $\gamma(n)$ will be described in the next subsection.

B. Average Number of Required Channels

The main responsibility of admission control is to estimate the network resources and make decisions based on the needs of users and available network resources. The number of CR free channels is one of the main network resources in CRSNs that should be estimated in order to decide about the admission of data flows.

In order to send sensors data toward the sink node, some CR channels are required. The number of these required channels depends on the system state, routing protocol and network topology (contending node number). The system state represents which sensors are sending their information toward the sink node. We consider the steady state behavior of routing protocol. In this way, a node selects one of the next hop nodes with a certain probability which does not change rapidly over time [12].

Therefore, for each sensor node, there are several possible routes toward the sink node. In order to decide about the admission of data flows in the network optimally, the optimal number of required channels should be estimated so that minimizing the
data packet collision. Assume there are \(K_i\) \((0 \leq i \leq N_S)\) possible routes between the sensor node \(i\) and the sink node. The sensor \(i\) uses its possible route \(d\) with the probability of \(P_{id}\). Therefore, there are \(\prod_{i=1}^{N_S} (K_i)^{n_i}\) possible combinations of routes for the data flows of admitted sensor nodes at each network state. Each possible combination of routes of the network state forms a network sub-graph. At each considered network sub-graph, the nodes have different number of contending nodes in the transmission of data packets to the sink node. In order to decrease the data packet collision, the optimal number of required channels at each possible combination of routes can be determined according to the maximum number of contending nodes of the nodes in the considered sub-graph. The problem of finding the optimal required number of channels at each possible combination of routes can be modeled by graph coloring approach. According to vertex coloring, different colors are assigned to each two adjacent vertex of the graph [13]. Each color label is equivalent to a CR free channel. The minimum number of required colors at each possible combination of routes can be considered as the minimum number of required channels.

Assume the minimum number of required channels at each possible routes configuration is considered as \(\Omega(l_1, n_1, l_2, n_2, \ldots, l_{N_S}, n_{N_S})\) where the \(l_1, l_2, \ldots, l_{N_S}\) are the selected route indexes of sensor 1, sensor 2, ..., sensor \(N_s\), respectively and also the \(n_1 \in \{0, 1\}, n = 1, 2, \ldots, N_S\) is the admission state of the sensor \(i\) which is described before. The value of the product \(l_i n_i\) will be zero when sensor \(b\) is not admitted and will be \(l_i n_i\) when sensor \(b\) is admitted. The notation of \(l_b\) is considered for the product \(l_b n_b\).

According to these definitions, the optimal average required number of channels at each state \(\gamma(n)\) can be calculated by Equ. 2.

\[
\gamma(n) = \sum_{l_1 \in \{0, 1\}} \sum_{l_2} \cdots \sum_{l_{N_S}} \left[\left(\prod_{i=1}^{N_S} (P_{l_i n_i})^{n_i}\right) \Omega(l_1, l_2, \ldots, l_{N_S})\right]
\]

(2)

The value of \(\Omega(l_1, l_2, \ldots, l_{N_S})\) is calculated by the minimum number of colors required for the network graph when the sensors 1, 2, ..., \(N_S\) are sending data packets in their \(l_1, l_2, \ldots, l_{N_S}\) routes toward the sink. Therefore, the \(\gamma(n)\) is the function of network state.

C. Action Space

At each decision epoch, an action \(a\) is selected as the result of the admission control decision for the next epoch. The action \(a\) at decision epoch \(t\) can be defined as \(a(t) = [a_1(t), a_2(t), \ldots, a_N(t)]\). The \(a_i(t) = 1\) represents the sensor \(i\) is admitted for sending data flow at decision epoch \(t\) and the \(a_i(t) = 0\) represents the rejection decision about this flow. Hence, the action space \(A\) can be defined as

\[
A = \{a \in \{0, 1\}, 0 \leq i \leq N_S, \sum_{i=1}^{N_S} a_i \leq 1\}
\]

(3)

The \(a = \{0, 0, \ldots, 0\}\) means that no data flow is admitted. At each decision epoch, the admission control mechanism decides about the admission of the sensors, sending request and at most admits one of the requesting sensors data flow. For each state, a subset of the action set \(A\) is valid; thus an action space for each state \(s \in S\) can be defined as

\[
A_s = \{a \in A: s = [n, q], [n + a, q] \in S\}
\]

(4)

D. State Transition

Assuming the states \(s = [n, q]\) and \(x = [n', q']\), the transition probability \(P_{sx}(a)\) is the probability of transition from state \(s\) to state \(x\) by selecting the action \(a\). There are some kinds of events in this admission control mechanism; (I) PU arrival to a channel that is free of CR user, (II) PU arrival to a channel that is using by a CR user and the CR user leaves the channel, (III) PU departure from a channel, and (IV) CR user arrival. When a PU departs from related channel, there is at least a CR user request in the queue to use this free channel. The event rates of the mentioned events are \(\sum_{i=1}^{N_S} r_a d_i (CH - \gamma_{n_i} - n_{s_i})\), \(\sum_{i=1}^{N_S} r_a (1 - d_i (CH - \gamma_{n_i} - n_{s_i}))\), \(\sum_{i=1}^{N_S} q_i d_i\), \(\sum_{i=1}^{N_S} a_i r_d (1 - d_i (CH - \gamma_{n_i} - n_{s_i}))\), respectively, where the function \(d_i\) can be defined as follows

\[
d_i = \begin{cases} 1 & i \geq 0 \\ 0 & i < 0 \end{cases}
\]

These events are independent Poisson processes, thus sum of these events follows the Poisson process too [14]. The total event rate of this system is the sum of event rates of the events (I), (II), (III) and (IV). Therefore, the inter-event time of this model is the reverse of total event rate. This inter-event time can be defined as the expected sojourn time of the SMDP. The sojourn time is the average time after action \(a\) is selected in current state \(s\) until the next decision epoch \(\tau_s(a)\).

\[
\tau_s(a) = \left(\sum_{i=1}^{N_S} r_a + \sum_{i=1}^{N_S} q_i r_d + \sum_{i=1}^{N_S} a_i r_d\right)^{-1}
\]

(4)

The transition probabilities can be derived using the decomposition property of the Poisson process. The transition probabilities between the states of this system can be determined as

\[
P_s(a) = \begin{cases} \rho_s \delta(CH - \gamma_{n_i} - q_{s_i})(a), & i = s + PU \\ q_i r_s r_t(a), & i = s - PU \\ \rho_s (1 - \delta(CH - \gamma_{n_i} - q_{s_i}) r_t(a), & i = s + PU - \psi(CR) \\ q_i r_s \delta(CH - \gamma_{n_i} - q_{s_i}) r_t(a), & i = s + CR \\ 0 & otherwise \end{cases}
\]

The \(s + PU\) and \(s - PU\) are the arrival and departure of a PU, respectively that are equivalent to \(s + [0, 1]\) and \(s - [0, 1]\), respectively. Also the \(s + CR\) and \(s - CR\) are equivalent to \(s + [1, 0]\) and \(s - [1, 0]\), respectively. The \(\psi(CR)\) is the representative of the worthless CR user who is transmitting data packets toward the sink node. The worth of CR users is determined based on their weight. According to this admission control mechanism, when a PU starts using its related channel while there is no free channel for
CR users, the most worthless CR user leaves using CR channel and stops sending data.

E. Policy and Reward Function

A policy \( \pi \) is a function that maps state space to acceptable action space. For each state \( s \in S \), an action is chosen according to policy \( \pi \). The \( \Pi \) is the acceptable policy space. The reward function \( R(s,a) \) is the average reward obtained from the network in current state \( s \) after the action \( a \) is selected until the next decision epoch. The reward function is the reward earned by the weight of new admitted CR user at each decision epoch. This function is defined as the sum of the weights of admitted flows to send to the sink node that can be defined as:

\[
R(s, a) = \sum_{i=1}^{N_S} a_i \omega_i. \tag{6}
\]

The average reward is considered as a performance measure. Inspiring from [14], the average reward function for \( \forall \pi \in \Pi \) is defined as:

\[
J_\pi(s_0) = \lim_{T \to \infty} \frac{1}{T} \int_0^T R(s(t), a(t)) dt \tag{7}
\]

where the \( s_0 \) is the first state that SMDP is started from and \( E(\cdot) \) is the expectation function. The purpose is to find the optimal policy \( \pi^* \in \Pi \) that maximizes the average reward for all initial states. On the other hand, the aim is to find the best policy that maximizes the average value of sent information via the admitted sensors.

F. Value Iteration Algorithm as a Solution of the SMDP

The suboptimal policy \( \pi \) can be obtained by the value iteration algorithm. The steps of the value iteration algorithm are as follows [12]:

1. Initialization: \( n = 1 \), choose a number \( \theta \) in the range of \( [0, \max r_a(\pi)] \) and \( \forall s \in S \) choose \( V_n(s) \) in the range of \( [0, \max R(s,a)] \).

2. \( \forall s \in S \) compute the function \( V_n(s) \) from Equ. 8 and obtain the stationary policy \( \pi(n) \) which is the maximum right hand side of Equ. 8. The \( V_n(s) \) function is the maximum obtained reward per time resulted by selection of an action \( a \) from action space in \( n \)th step of algorithm that is the function of \( V_{n-1}(s) \).

\[
V_n(s) = \max_{a \in A(s)} \frac{R(s, a)}{r_a(\pi)} + \theta \sum_{s' \in S} p_{s,s'}(\pi)V_{n-1}(s') + (1 - \theta r_a(\pi))V_{n-1}(s) \tag{8}
\]

3. The algorithm is stopped with stationary policy \( \pi(n) \) when \( 0 \leq M_n - m_n \leq \epsilon \), otherwise go to next step. The \( \epsilon \) is the specified accuracy number and the values of \( M_n \) and \( m_n \) are computed as follows:

\[
M_n = \max_{s \in S} \{ V_n(s) - V_{n-1}(s) \} \tag{9}
\]

\[
m_n = \min_{s \in S} \{ V_n(s) - V_{n-1}(s) \} \tag{10}
\]

4. \( n = n + 1 \) and go to step 2.

The steps of this algorithm are iterated for finitely much number of iterations. The value of \( \delta \) is recommended to set as \( \max_{a,s} \tau_s(a) \).

IV. EXPERIMENTAL RESULTS

In this section, the performance of the proposed mechanism is evaluated through CogNS that is a simulation framework based on NS-2 [15] for cognitive radio networks [16]. A CR sensor network is placed in a 50m × 50m field. The number of the PUs and frequency channels is taken as 6. It is assumed each PU individually has the license of using related frequency channel. The values of \( N_p, N_c \) and \( N_{CR} \) are set as 8, 3 and 7, respectively. The sensing time and operating time are considered as 0.01 and 0.6 sec, respectively. The default values of PUs arrival and departure rates are considered as 1; these two rates are changed for different experiments. The packet size is considered 100 bytes. The simulation time is 200 second. The value of termination parameter in value iteration method (\( \epsilon \)) is considered as 0.02. Each experiment is run five times, and the results are averaged.

The proposed admission control mechanism is evaluated in this section by several experiments in different PU activity settings. The PU activity \( (r_a, r_d) \) is determined based on the length of ON and OFF periods of PU transmissions. When the PU arrival rate \( (r_a) \) is greater than the PU departure rate \( (r_d) \), this state is considered as a \( \chi \) high PU activity state. Furthermore, when the PU arrival rate is equal to PU departure rate, this state is considered as a \( \chi \) medium PU activity state. Also, when the PU arrival rate is equal to PU departure rate, this state is considered as a \( \chi \) low PU activity state. According to these definitions, the PU activities (3,1) and (5,1) belong to the low PU activity state, the PU activities (1,1), (3,3) and (5,5) belong to the medium PU activity state, and the PU activities (1,2), (1,3), (1,4), (1,5) and (1,6) belong to the high PU activity state.

In this section, the performance of the introduced mechanism (referred as SMDP-based) is evaluated and compared with the proposed mechanism in [3] (referred as Threshold-based) and the network without applying the admission control (referred as complete sharing). The Fig. 1, Fig. 2 and Fig. 3 illustrate the packet loss probability, jitter and end to end delay, respectively, for three scenarios, i.e., complete sharing, the network with SMDP-based admission control mechanism and the network with Threshold-based admission control mechanism, with regard to different PU activities. This admission control estimates the average required channels of the flows. According to this estimation, more valuable flows are admitted to send data toward the sink. As depicted in these figures, the packet loss ratio of the network is reduced with the proposed SMDP-based admission control, especially in high PU activities, i.e., (1,3) and (1,5). As depicted in these figures, the SMDP-based admission control mechanism reduces the number of flows according to the decisions of the proposed
admission control leads to reduce the average jitter of data packets as depicted in Fig. 2. Furthermore, this mechanism reduces the packet end to end delay as illustrated in Fig. 3.

Fig. 4 represents the average reward earned by the optimal policy that is a decision maker, in the different states of the network. This figure illustrates the average reward per second in the networks with different channel numbers with regard to different PU activities. The channel numbers varies from 3 to 7. The existence of the more channels in the network leads to admit the more number of data flows and earn more reward. In the low PU activities the network earns highest reward and also in the high PU activities the network earns lowest reward due to the more active PUs. The highest reward is earned in PU activity (5,1) and channel number 7.

Fig. 5 depicts the throughput of the networks with different channel numbers for different PU activities. The numbers of channel varies from 3 to 7. As illustrated in this figure, network throughput decreases with the increase of the PU entrance rate or the decrease in the number of channels. The highest throughput is obtained in PU activity (1,1) and channel number 7.

V. CONCLUSION

In this study, a suboptimal optimal connection admission control (CAC) mechanism is proposed in order to provide QoS of the CR users in cognitive radio sensor networks. This mechanism is modeled as a semi Markov decision process (SMDP) and a suboptimal policy is obtained by a value iteration method. This proposed mechanism decreases the jitter, end to end delay and packet loss ratio of the packets in the network. The performance of the CAC is evaluated by NS-2 based simulation. The simulation results represent that the proposed mechanism outperforms the previous proposed admission control mechanism in CRSNs. Due to the requirements of CRSNs, the end to end delay and power constraints can be added to this SMDP model as future work.

ACKNOWLEDGMENT

The authors would like to thank Iran Telecommunication Research Center (ITRC) for supporting this research.

REFERENCES


Elahe S. Hosseini received her BS and MS degrees in computer engineering from the Dr. Shariati College, Tehran, Iran, in 2006 and the Sharif University of Technology (SUT), Tehran, Iran, in 2008, respectively. During her PhD studies in 2012, she spent six months as a visiting researcher at Koc University, Istanbul, Turkey. She is currently a PhD candidate in the Computer Engineering Department at the Iran University of Science and Technology (IUST) in the area of cognitive radio networks.

Reza Berangi received his B.Sc. degree in telecommunication and his M.Sc. degree in electronic engineering from the Iran University of Science and Technology (IUST), Tehran, Iran, in 1985 and 1989, respectively. He received his Ph.D. degree in mobile communications from Victoria University, Melbourne, Australia, in 1998. From 1997 to 1999, he was with the Mobile Communications and Signal Processing Group at VUT, working on smart antenna mobile base stations.
Fig 1: Average packet loss probability in different PU activities in the network with complete sharing, the networks with SMDP-based and Threshold-based admission control mechanisms

Fig 2: Average jitter in different PU activities in the network with complete sharing, the networks with SMDP-based and Threshold-based admission control mechanisms

Fig 3: Average packet end-to-end Delay in different PU activities in the network with complete sharing, the networks with SMDP-based and Threshold-based admission control mechanisms

Fig 4: Average reward per second in different channel numbers and different PU activities

Fig 5: Average throughput in different channel numbers and different PU activities
This Page intentionally left blank.
QoS-aware web service composition using Gray Wolf Optimizer

Meysam Karimi
Department of computer science
Kashan University
Kashan, Iran
Meysam.Karimi84@gmail.com

Seyed Morteza Babamir
Department of computer science
Kashan University
Kashan, Iran
Babamir@kashanu.ac.ir

Received: November 4, 2016- Accepted: March 5, 2017

Abstract— In a service-oriented application, an integrated model of web services is composed of multiple abstract tasks. Each abstract task denotes a certain functionality that could be executed by a number of candidate web services with different qualities. The selection of a web service among candidates for execution of each task that is led to an optimal composition of selected web services is a NP-hard problem. In this paper, we adapt the Gray Wolf Optimizer (GWO) algorithm for selection of candidate web services whose composition is optimal. To evaluate the effectiveness of the proposed method, four quality parameters, response time, reliability, availability, and cost of web services are considered and the derived results are compared with several Particle Swarm Optimization (PSO) methods. The proposed method was executed in from 100 to 1000 times and the results showed that a better optimal rate (between 0.2 and 0.4) compared with PSO.

Keywords- Optimal web service composition; Gray Wolf Optimizer algorithm; Particle Swarm Optimization; Service oriented; Quality of service.

I. INTRODUCTION

In recent years, web services as computational models were developed quickly and played significant roles in e-commerce and web-based services. Therefore, the use of convenient and fast web service with atomic functionality has increased. However, for an application consisting of tasks, a combination of web services is used to execute the tasks where each task (called abstract task) is meant for a specific function. For each task, there are a number of candidate web services with the same functionality but with different quality characteristics. An optimal solution for execution of an application is a set of selected web services whose combination is the most suitable combination for the application. Since optimal values of quality parameters are not included just in a candidate web service and are found in different candidates, the selection of a candidate web service for execution of a task of the application is difficult. Furthermore, there may be conflict between some quality parameters. Lower cost and faster response time are always desired; however, they are in conflict with each other because a web service with faster (more optimal) response time demands more (less optimal) cost. Hence, it is clear that the web service composition is a combinatorial optimization problem. It is worth noting that the quality parameters play an important role in identifying the best combination of services at runtime. Finding the optimal solutions for
web services composition with conflicting quality parameters is a complex problem that cannot be solved in a polynomial time (NP-Hard).

Generally speaking, the QoS-aware (Quality of Service-aware) web services composition is resolved using intelligent computational methods [1-3]. Methods used the PSO algorithm exhibit better results compared with genetic-based methods [4]. In this study, we adapted the Gray Wolf Optimizer (GWO) algorithm to resolve the QoS-aware web services composition. The proposed method was compared with standard PSO (Particle Swarm Optimization) [5], IDPSO (Improved Discrete PSO) [6] and QIPSO (Quantum-Inspired PSO) [7]. The results showed that the proposed method is more effective. We've already had experience using the adapted GWO for optimization where the GWO results were compared with those of other optimization algorithms [8].

This paper is organized as follows. Section 2 addresses the related literature. The GWO algorithm is described in Section 3. In Section 4, we explain the proposed method and in Section 5, present the results. Finally, Section 6 deals with concludes.

II. RELATED WORKS

A number of studies have been carried out for the QoS-aware web service composition problem. Most solutions are based on the PSO algorithm and exhibit better coverage compared with the genetic optimization algorithm. However, there is still room for optimization of the QoS-aware web service composition [6]. PSO is a population-based evolutionary algorithm in which each particle has a position and velocity and the population of particles saves its best local and global position. Each particle improves its position based on the value of: (1) its position, (2) the best local position (pbest) and (3) the best global position (gbest). Each particle has a D-dimensional vector in which $D$ represents dimension of the space that the particle wants to search.

Kang et al [5] used PSO to solve the problem of QoS-aware web service composition composed of the following stages. They noted that the results in terms of coverage and execution speed are superior to the genetic algorithm.

I) Reduction of the multi-objective optimization problem to a single-objective one,
II) Initialization of the particles and adjust the parameters of the algorithm,

\[
\text{for each particle} \{ \\
\text{III) Computation of the fitness value of the particle position as a candidate for the composition,} \\
\text{IV) Comparison of the fitness value of current position of the particle with pbest of the particle and replacement of the pbest by the fitness value if the value is more,} \\
\text{V) Comparison of the pbest value of the particle with the gbest and replacement of the gbest by the pbest if the pbest value is more,} \\
\text{VI) Calculation of velocity and position of the particle using the PSO formulas,} \\
\}
\]

Zhao et al. [6] used IDPSO to address discrete QoS-aware web service composition. They modified the PSO position and velocity formulas to resolve the QoS-aware web services composition and showed that the quality of the service based on the composition obtained by IDPSO is higher than PSO. We compared our results with IDPOS.

QIPSO [9] was created by the integration of quantum display of problem space and PSO trying to improve the ability of the PSO algorithm. Jatush and Gangazaran [7] first reduced the QoS-aware web service composition problem to the single-objective optimization and then resolved it through QIPSO. QIPSO contains three basic parts: (1) quantum measurement, (2) quantum interference, and (3) quantum flight.

Quantum measurement is a function to extract binary particles from quantum particles. Consequently, the quantum particles can be transformed to binary vectors in the problem space. Quantum interference is a function increasing the composition optimization and decreasing the probability of suboptimal composition. The main purpose of the quantum interference is that the state of each qubit tends towards the optimum composition (solution). A qubit in quantum computing or quantum bit is a basic unit like a bit in the classical computing. Quantum flight is a function allowing a quantum moves from its current position to its next one to enhance the capacity of the search space. A new solution uses standard phrases forming the next position of the particle in the PSO algorithm. It was shown that QIPSO is more effective than PSO and IDPSO.

III. GRAY WOLF OPTIMIZER

Gray Wolf Optimizer (GWO) [10] is a population-based meta-heuristic algorithm that simulates leadership structure and hunting mechanism of gray wolves in nature. Gray wolves prefer to live in a group of five to twelve in form of a hierarchical society consisting of four levels: Alpha, Beta, Delta, and Omega.

The Alpha wolves (male or female) are leaders and responsible for deciding on time of hunting, sleeping, waking, and so on. The rest of the wolves in the group are forced to obey the order of Alphas. Alphas prevail over other levels and all their orders must be followed by members of the group.

The Beta wolves (male or female) are subalterns of Alphas and help Alphas in decision-making. They are the best alternatives to the Alphas at the time of death or aging.

The Delta wolves obey Alphas and Betas, but are superior to the Omegas. Omegas are considered as devotees and obey all wolves of their higher levels. They are the last ones allowed to eat.

GWO simulates hunting of gray wolves where the hunting process is divided into three phases: (1) to chase and surround a prey, (2) harass the prey, and (3)
attack the prey. For mathematical modeling of the problem, the best solution is considered as Alpha (α). Similarly, the second and third best solutions are considered as Beta (β) and Delta (δ). Remaining candidate solutions are considered as Omega (ω). Hunting (optimization) in GWO is guided by Alphas, Betas, and Deltas while Omegas follow them. Eqs. 1 and 2 are used to surround a prey [10]:

\[
\vec{D} = |\vec{C}_{\alpha} \vec{X}_{\alpha}(t) - \vec{X}(t)|
\]

\[
\vec{X}(t + 1) = \vec{X}(t) - \vec{A} \cdot \vec{D}
\]

As stated above, wolves should surround the prey first. To this end, the distance of each wolf from the prey is calculated according to Eq. 1 and the next position of the wolf is calculated according to Eq. 2 where \(t\) represents the current run and vectors \(\vec{A}\) and \(\vec{C}\) are coefficient vectors for distance and prey respectively. \(\vec{X}_p\) and \(\vec{X}\) are the prey position vector and position of the gray wolf, respectively. Vectors \(\vec{A}\) and \(\vec{C}\) are calculated according to Eqs. 3 and 4 [10]:

\[
\vec{A} = 2\alpha \vec{r}_1 - \vec{a}
\]

\[
\vec{C} = 2\vec{r}_2
\]

Elements of the vector \(\vec{a}\) linearly decrease from 2 to zero during the execution of algorithm. Vectors \(r_1\) and \(r_2\) contain random values in interval [0, 1].

The wolves chase the prey based on positions of Alphas, Betas, and Deltas. Wolves get away from each other for searching (called divergence) but get close to each other to attack (called convergence). To model the divergence, values of the \(\vec{A}\) vector are greater than 1 or less than -1. It forces the search agent to diverge and perhaps find a better prey. This practice facilitates exploration and allows GWO to perform a complete search. Another part of GWO, which facilitates exploration, is the \(\vec{C}\) vector whose values are random values in [0, 1] denoting weight of prey in Eq. 1. It causes the behavior that is more random during hunting leading to find a proper prey; this avoids the local optimization. Note that the \(\vec{C}\) vector decreases non-linearly. Moreover, the \(\vec{C}\) vector can be interpreted as natural obstacles in the path of the wolves in hunting and preventing them from rapidly reaching the prey. The wolves are able to recognize and surround a prey position. The hunting of prey usually is guided by Alphas and Betas and Deltas participate in hunting in some cases. However, for the optimization problem, there is no information about the exact position of the prey. To model the behavior of hunting, Eqs. 5-7 are used [10]:

\[
\vec{D}_i = |\vec{C}_{\alpha} \vec{X}_i - \vec{X}_p|, \vec{D}_j = |\vec{C}_{\beta} \vec{X}_j - \vec{X}_p|, \vec{D}_k = |\vec{C}_{\delta} \vec{X}_k - \vec{X}_p|
\]

\[
\vec{X}_i = \vec{X}_i - \vec{A}_i (\vec{D}_i), \vec{X}_j = \vec{X}_j - \vec{A}_j (\vec{D}_j), \vec{X}_k = \vec{X}_k - \vec{A}_k (\vec{D}_k)
\]

\[
\vec{X}(t + 1) = \frac{\vec{X}_i + \vec{X}_j + \vec{X}_k}{3}
\]

Since there is not the precise estimate of the actual location of the prey, the distance of each wolf from the best positions of Alpha, Beta, and Delta is calculated using Eq. 5. The next position of Alpha, Beta, and Delta is calculated using Eq. 6. Using Eq. 7, next wolf position is calculated regarding the average position of Alpha, Beta, and Delta.

IV. PROPOSED METHOD

In this section, an accurate description of the problem is defined and then quality parameters of the QoS-aware web service composition are described. Afterwards, three steps are taken to solve the optimization problem using GWO.

A. Problem Description

An abstract description of a workflow is defined as a composition of abstract services indicated by \(A = (A_1, A_2, \ldots, A_n)\) where \(A_i\) is an abstract service. Suppose for each abstract service, there are some candidate concrete services that are able to perform the abstract service with different qualities. Concrete candidate services for abstract service \(A_i\) is shown as \(C_{ij} = \{C_{i1}, C_{i2}, \ldots, C_{im}\}\) where \(C_{ij}\) is the \(j^{th}\) concrete candidate service for abstract service \(A_i\).

If quality attributes are response time, cost, availability, reliability, quality of the concrete service \(S\) is defined as:

\[
\text{QoS}(S) = (\text{Time}(S), \text{Cost}(S), \text{Availability}(S), \text{Reliability}(S))
\]

The goal is to obtain an optimal composition of candidate concrete services for services of a workflow so that the composed web services have the best QoS. Therefore for each abstract service, say \(A_i\), the goal is to find solution \(S_i = C_{ij}\). For a workflow consisting of abstract services \(A_1, A_2, \ldots, A_n\), an optimal composition consisting of candidate services \(C_{1j_1}, C_{2j_2}, \ldots, C_{nj_n}\) should be obtained with respect to minimizing response time and cost and maximizing reliability and availability. Notation \(C_{ij}\) indicates the selected concrete service for abstract service \(A_i\) is \(j_i\) where \(j_i\) denotes the \(j^{th}\) concrete service of the services are candidate for \(A_i\).

Depending upon the execution of concrete services in serial or in parallel, response time, cost, reliability, and availability are calculated according to Table 1 [11].

B. QoS-aware web service composition using GWO

To optimize QoS-aware web service composition using GWO, we should first determine: (1) the representation of wolves, (2) the initial population of wolves, (3) the fitness function to evaluate wolves and (4) the mechanism of updating wolves, positions at the end of each algorithm iteration.
C-1 Wolf representation

One of the most important steps in the GWO design is the representation of a solution (wolf). In a QoS-aware web service composition, a proper solution is shown by a vector with \( D \) dimensions (called \( D \)-dimensional vector) in which \( D \) is the number of abstract tasks of the workflow. Each element of the vector has a value (see Eqs. 5 to 7) indicating index of the concrete service selected from the candidate services.

Consider a workflow consisting of \( (n=5) \) abstract services, for instance, where \( n \) solutions (indicated by vectors \( x_1 \) to \( x_n \) in Fig. 1) were proposed. Each solution, indicated by \( x_i \), shows a composition of 5 candidate services. Vector \( x_1 \), for instance, indicates that concrete services 3, 1, 4, 11, and 6 are selected for abstract services 1 to 5.

\[ x_1 = \{3,4,11,6\}, \quad x_2 = \{110,68,63,400,100\}, \ldots \]

\[ x_n = \{40,36,92,57,102\} \quad \text{Fig. 1} \]

C-2 Initializing population

After representing solutions, a population of solutions should be initialized. Initially, \( n \) wolves (solutions) are randomly chosen for each abstract task from candidate services in a dataset. Each wolf consists of \( d \) values (for instance 5 values for the example stated above).

C-3 Fitness Function

A fitness function should be determined to measure wolf’s accuracy. For QoS-aware web services composition, wolf’s accuracy is measured by its service quality values and considering the importance (weight) of each service quality. To compute quality of web services, we use the relations stated in Table 1. Table 2 shows typical weights for quality services.

Availability and reliability are positive qualities while cost and response time are negative ones. While higher values are more desirable for positives, fewer values are sought for negatives. Because qualities values have different scales, they should be normalized. Eqs. 8 and 9 show normalization of positive and negative qualities, respectively [6].

### Table 1. Calculation of quality parameters in execution of serial or parallel services

<table>
<thead>
<tr>
<th>Quality parameter</th>
<th>Serial</th>
<th>Parallel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response Time</td>
<td>( \Sigma(T_i) )</td>
<td>( \max(T_i) )</td>
</tr>
<tr>
<td>Cost</td>
<td>( \Sigma(C_i) )</td>
<td>( \Sigma(C_i) )</td>
</tr>
<tr>
<td>Availability</td>
<td>( \Pi(A_i) )</td>
<td>( \Pi(A_i) )</td>
</tr>
<tr>
<td>Reliability</td>
<td>( \Pi(R_i) )</td>
<td>( \min(R_i) )</td>
</tr>
</tbody>
</table>

### Table 2. Typical weight (importance) for each quality

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Cost</th>
<th>Availability</th>
<th>Response time</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>0.1</td>
<td>0.2</td>
<td>0.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

\[
\frac{Q_i^{\max} - Q_i}{Q_i^{\max} - Q_i^{\min}} \quad \text{if } Q_i^{\max} - Q_i^{\min} \neq 0
\]

\[
\frac{Q_i^{\max} - Q_i^{\min}}{Q_i^{\max} - Q_i^{\min}} \quad \text{otherwise}
\]

According to Eq. 8, higher (more desirable) value for the positive quality \( Q_i \) leads to less value for the fraction; similarly, based on Eq. 9, less (more desirable) value for negative the quality \( Q_i \) does to less value for the fraction. Therefore, our optimization problem is a minimization problem.

Given that the Max. and Min. values of reliability and cost are according to Table 3 and reliability and cost values of a service are according to Table 4, normalized values are shown in Table 5.

As stated above, our optimization is a minimization problem; therefore, according to Eqs. 10 and 12, Service1 is more reliable than Service 2 but Service2 is less costly than Service1.

Fitness value of each dimension (indicating a concrete service) in a candidate solution (wolf) is calculated according to Eq. 14. For instance, fitness value of concrete Service 3 consisting of values of 100% for response time, 2.2% for availability, 90% for cost, and 89% for reliability are calculated as follows (for weights, see Table 2).

Table 6 shows Fitness values of services in vector \( \bar{x}_1 = \{3,4,11,6\} \) (see Fig. 1) as a solution for 5 abstract services of a workflow.

Eq. 15 shows the fitness value of solution \( \bar{x}_1 \); similarly, the fitness values of \( \bar{x}_2 \) to \( \bar{x}_n \) are calculated and the solution with the smallest value is selected as Alfa wolf and the smallest values greater than Alfa are selected as Beta and Delta wolves, respectively.

### Table 3. Min. and Max. values of reliability and cost

<table>
<thead>
<tr>
<th>Reliability(%)</th>
<th>Maximum</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost($)</td>
<td>Maximum</td>
<td>Minimum</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>20</td>
</tr>
</tbody>
</table>
of the approximate position of wolves in continuous space, update formulas for wolf position are applied.

The next position of a wolf (calculated using Eq. 7) is used as argument of the hyperbolic tangent function in Eq. 16 and the output of the function is compared with a random number between zero and one. If it is lower than the random value, it means that we do not need to change the concrete service; otherwise, we must replace the concrete service with a new one.

For example, suppose Alpha, Beta, and Delta are defined as follows:

\[ \mathbf{x}_* = \{500,12,139,41,7\} \quad \mathbf{x}_* = \{17,29,219,76,900\} \quad \mathbf{x}_* = \{2028,1101,739,606,84\} \]

To update solution (wolf) \( \mathbf{x}_* = \{3,1,4,11,6\} \), distance of each dimension of the concrete service from the corresponding dimensions of Alpha, Beta, and Delta is calculated and then a new approximate position is calculated according to Alpha, Beta, and Delta separately.

\[ X_{new} = \begin{cases} 1, & \text{if } \left| \text{Tanh}\left((X(t+1))_a\right) > U(0,1) \right| \\ 0, & \text{otherwise} \end{cases} \]

Moreover, the mean of these positions is used in the transition function. Afterwards, output of this function determines whether this service should be replaced or not. The following example shows the calculations stated above.

\[ \mathbf{D}_{a1} = \mathbf{C}_1 \mathbf{x}_a - \mathbf{x}_* \Rightarrow 0.58 \times \mathbf{0.92} - 0.83 = 29.10 \]
\[ \mathbf{D}_{b1} = \mathbf{C}_2 \mathbf{x}_b - \mathbf{x}_* \Rightarrow 0.58 \times \mathbf{0.88} - 0.83 = 31.45 \]
\[ \mathbf{D}_{d1} = \mathbf{C}_3 \mathbf{x}_d - \mathbf{x}_* \Rightarrow 0.58 \times \mathbf{0.95} - 0.83 = 27.34 \]
\[ \mathbf{x}_1 = \mathbf{x}_a - \mathbf{A}_1 \mathbf{D}_{a1} \Rightarrow 0.92 - (-1.82) \times 2.10 = 144.95 \]
\[ \mathbf{x}_2 = \mathbf{x}_b - \mathbf{A}_2 \mathbf{D}_{b1} \Rightarrow 0.88 - (-0.89) \times 3.45 = 116.22 \]
\[ \mathbf{x}_3 = \mathbf{x}_d - \mathbf{A}_3 \mathbf{D}_{d1} \Rightarrow 0.95 - (0.78) \times 27.34 = 73.77 \]
\[ X_{1}(t+1) = \frac{\mathbf{x}_1 + 2 \mathbf{x}_2 + \mathbf{x}_3}{3} \]
\[ \Rightarrow 144.95 + 116.22 + 73.77 = 111.64 \]

\[ \text{TransformValue} = \begin{cases} 1, & \text{if } \left| \text{Tanh}\left((X(t+1))_a\right) > U(0,1) \right| \\ 0, & \text{otherwise} \end{cases} \]

C-4 Update wolf position

In GWO, wolves need to update their position at the end of each algorithm iteration according to Alpha, Beta, and Delta (which are the wolves with the best fitness values in the population), to get closer to the prey. The classic GWO is not appropriate for solving discrete problem and since the web service composition has a discrete space (each dimension of web services composition is a representation of one dimension of a concrete service and cannot accept continuous values), we should justify the basic GWO to a discrete problem. Kennedy and Eberhart [4] used a sigmoid method to convert continues problems into discrete ones. Mirjalili et al. [12] described different ways of transforming continuous problems to discrete ones. Leading solutions were considered in this study and the results showed that the hyperbolic tangent function is more suitable for our problem. Therefore, after using GWO equations (Eqs. 5-7) and calculation

**Table 4. An example of quality values of 2 services**

<table>
<thead>
<tr>
<th>Service 1</th>
<th>Reliability</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80</td>
<td>60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Service 2</th>
<th>Reliability</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>70</td>
<td>40</td>
</tr>
</tbody>
</table>

**Table 5. Normalizing quality values**

<table>
<thead>
<tr>
<th>Concrete service</th>
<th>Reliability(%)</th>
<th>Cost(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service 1</td>
<td>100-80=20/80=0.25</td>
<td>60-20=40/80=0.67</td>
</tr>
<tr>
<td>Service 2</td>
<td>100-70=30/80=0.37</td>
<td>40-20=20/80=0.33</td>
</tr>
</tbody>
</table>

**Table 6. Fitness values of the services in vector x**

<table>
<thead>
<tr>
<th>Service</th>
<th>Response time (%)</th>
<th>Availability (%)</th>
<th>Cost (%)</th>
<th>Reliability (%)</th>
<th>Fitness value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>100</td>
<td>2.2</td>
<td>90</td>
<td>89</td>
<td>76</td>
</tr>
<tr>
<td>1</td>
<td>85</td>
<td>1</td>
<td>100</td>
<td>67.5</td>
<td>34.27</td>
</tr>
<tr>
<td>4</td>
<td>83</td>
<td>23</td>
<td>6</td>
<td>65.18</td>
<td>77.73</td>
</tr>
<tr>
<td>11</td>
<td>73</td>
<td>3.7</td>
<td>12</td>
<td>54.54</td>
<td>58.05</td>
</tr>
<tr>
<td>6</td>
<td>46</td>
<td>10</td>
<td>39</td>
<td>89</td>
<td>60.11</td>
</tr>
</tbody>
</table>

**Update wolf position**

**Table 5. Normalizing quality values**

<table>
<thead>
<tr>
<th>Concrete service</th>
<th>Reliability(%)</th>
<th>Cost(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service 1</td>
<td>100-80=20/80=0.25</td>
<td>60-20=40/80=0.67</td>
</tr>
<tr>
<td>Service 2</td>
<td>100-70=30/80=0.37</td>
<td>40-20=20/80=0.33</td>
</tr>
</tbody>
</table>

**FitnessValue**=(\( \sum_{i=1}^{D} \text{weight}(quality(i)) \)*quality(i))

=\((0.4*1)+(0.02*0.2)+(0.9*0.1)+(0.89*0.3)\)=76%

\(X_{new} = \begin{cases} 1, & \text{if } \left| \text{Tanh}\left((X(t+1))_a\right) > U(0,1) \right| \\ 0, & \text{otherwise} \end{cases} \)

Because the value of the transition function is 1, this concrete service is replaced by one of randomly selected candidate services of this abstract service. The presented calculations were done for the first dimension of a candidate solution (wolf). Similarly,
this process should be done for all dimensions of a wolf. Pseudo code of proposed method is presented in Fig. 2.

```csharp
InitializePopulation();
currentIteration=0;
While(currentIteration < maxIteration){
    GetQualityParameters(population);
    updatePopulation(population, Alpha, Beta, Delta)
    currentIteration++;
}
InitializePopulation()
Foreach dimension in Dimension
    SelectRandomService(Repository);
}
updatePopulation(population, Alpha, Beta, Delta){
    foreach wolf in population
        GetNewConcreteServiceFromRepository();
}
```

Fig. 2. Pseudo code of the proposed method

V. EVALUATION OF THE PROPOSED METHOD

We implemented the QoS-aware web service composition using GWO in C# programming language and tested using PC with Intel® Core (TM) i5, 2.6 GHz and RAM of 8 GB. The optimal rate of the proposed method was compared with PSO [5], IDPSO [6] and QIPSO [7] using the QWS dataset [13] containing 2507 real web services. The optimality rate was calculated using Eq. 17 [14].

$$\text{Optimality Rate} = \frac{\text{Optimal Solution}}{\text{Initial Solution}}$$ (17)

The Initial Solution is the best solution at the end of the first iteration of the algorithm and the Optimal solution is the best solution after convergence of the algorithm. Table 7, for instance, shows initial and optimal solutions of algorithms A and B. According to Eq. 17, the optimal rate of algorithms A and B are 0.33 and 0.5, respectively. Since the minimum value of the optimal rate is desired, algorithm A outperforms algorithm B. To compare our proposed method with other methods, the optimal rate is used (Fig. 3). As Fig. 3 shows, the optimal rates of QIPSO and IDPSO are better than PSO and QIPSO outperforms IDPSO when the number of iterations increases. However, the optimal rate of the proposed method shows it outperforms other methods.

A suitable algorithm is the one that produces effective results independently from the number of the algorithm iterations; GWO enjoys such feature. In this study, we run the proposed algorithm 40 times with an arbitrary number of iterations in each execution. Table 8 shows convergence of the algorithm for 10 services.

VI. CONCLUSION AND FUTURE WORKS

In this study, the effective use of GWO for QoS-aware web service composition was investigated. To find an optimal composition of solutions in a discrete space, we modified the basic GWO.

To evaluate the effectiveness of the proposed method, we thought of quality parameters: response time, reliability, availability, and cost for each web service. By comparing results of the proposed method with several variations of PSO, it was shown the proposed method outperforms the various PSO-based methods.

We showed that the GWO was a suitable algorithm to produce effective results independently from the number of the algorithm iterations. In this study to produce effective results and not using non-optimal results, we run the proposed algorithm 40 times with an arbitrary number of iterations in each run.

The disadvantage of the proposed method, however, is that if a web service has the best fitness value it will be selected as the suggested solution; while there may be several similar solutions with lower fitness values but with more user-friendly candidates. In this case, these solutions would stay away from users. Therefore, as future work, we plan to solve QoS-aware web-service composition problem using the Pareto front concept without transforming it to the single-objective optimization problem. This would produce good non-dominated results and user would be free to decide between several suggested solutions.

As a future work, we may use Analytical Hierarchical Process (AHP) when we reach to a set of optimal solutions instead of single one. AHP is used when we should choose a solution from a set of alternatives. A solution is chosen from alternatives by considering some criteria influencing on the solutions.

Table 7. Initial and optimal solutions of 2 algorithms

<table>
<thead>
<tr>
<th>Alg</th>
<th>Initial Solution</th>
<th>Optimal Solution</th>
<th>Init. S./ Opt. S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.6</td>
<td>0.2</td>
<td>0.2/0.6=33</td>
</tr>
<tr>
<td>B</td>
<td>0.4</td>
<td>0.2</td>
<td>0.2/0.4=55</td>
</tr>
</tbody>
</table>
Fig. 3- The optimality rate of PSO, IDPSO, GWO, and QIPSO

Table 8. Convergence of the proposed algorithm
OG: #Optimal Generations, RI: #Run Iterations

<table>
<thead>
<tr>
<th>OG</th>
<th>RI</th>
<th>OG</th>
<th>RI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32</td>
<td>600</td>
<td>401</td>
</tr>
<tr>
<td>2</td>
<td>131</td>
<td>700</td>
<td>325</td>
</tr>
<tr>
<td>3</td>
<td>242</td>
<td>800</td>
<td>597</td>
</tr>
<tr>
<td>4</td>
<td>211</td>
<td>900</td>
<td>702</td>
</tr>
<tr>
<td>5</td>
<td>273</td>
<td>1000</td>
<td>654</td>
</tr>
</tbody>
</table>

REFERENCES


Seyed Morteza Babamir received BS degree in Software Engineering from Ferdowsi University of Mashhad and MSc and PhD degrees in Software Engineering from Tarbiat Modares University in 2002 and 2007 respectively. He was a researcher at Iran Aircraft Industries, in Tehran, Iran, from 1987 to 1993, head of Computer Center in University of Kashan, Kashan, Iran, from 1997 to 1999 and head of Computer Engineering Department at University of Kashan from 2002 to 2005. Since 2007, he has been an associate professor of Department of Computer Engineering at University of Kashan, Kashan, Iran. He authored one book in Software Testing, 4 book chapters, 20 journal papers and more than 50 international and national conference papers (http://ce.kashanu.ac.ir/babamir/Publication.htm). He is managing director of Soft Computing Journal published by supporting University of Kashan, Kashan, Iran.

Meysam Karimi received the B.E. degree in software engineering from the Tabarestan University, Chalus, Iran, in 2007, and M.Sc. degree in Software Engineering from University of Kashan, Kashan, Iran in 2016, respectively. In 2008, he joined the Lovin Information Technology Company; a software house digital repository under .Net technology, as a full-time Developer, and soon was promoted to the project manager in 2009. Since that time, he has been with the Lovin, where he was a technical manager, Scrum Master and part-time developer. In 2010, he joined Ministry of Education as a teacher and he has been teaching computer courses for students eagering to join an associate's degree. Also, he has held lecturing position at Azad University of Islamshahr, Tehran, Iran. He was the recipient of the best teacher of the Conservatory of Mostafa Khomeini in 2014, 2016, and 2017. His current research interests include software engineering, service oriented, cloud computing, distributed systems, evolutionary algorithms, and big data.
Inconsistency Repair to Improve the Alignment Results of Ontology Matchers

Bahareh Behkamal
Department of Computer Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
b.behkamal@mail.um.ac.ir

Mahmoud Naghibzadeh
Department of Computer Engineering
Ferdowsi University of Mashhad,
Mashhad, Iran
naghibzadeh@um.ac.ir

Received: October 22, 2016- Accepted: March 14, 2017

Abstract—Ontology inconsistency is one of the most important topics in the field of ontology matching. Until now many matchers are introduced but most of them suffer from inconsistencies. Many of the ontology matching tools have severe problems with respect to the quality of matching results and therefore the results of matching process is not adequate. In this paper, we focus on this topic and present a new method to produce better results from the matching process. The major novelty of this paper is in detecting the inconsistencies in ontologies before starting the matching process. In this phase, many problems caused by ontology diversity are resolved. Besides, some new patterns and inconsistencies in ontologies are detected and then refactoring operations are applied on them. At the end, one of the well-known matchers in OAEI is selected to evaluate our work. Experimental results show that the transformed ontologies are more efficient than original unrepaired ones with respect to the standard evaluation measures.

Keywords- Ontology matching; Alignment; Inconsistency; Refactoring; Pattern detection

I. INTRODUCTION

The vast progress of data and communication on the web has caused a huge amount of diversity in information. The problem of managing heterogeneity in various information resources is increasing. Until now many solutions have been proposed to facilitate this problem, and specifically, to automate integration of distributed data resources. Among them, semantic technologies have attracted particular attention. One of the best semantic technologies in this field is ontology matching. Ontology matching is a technique that takes the ontologies as an input and extracts the alignments as an output. The alignment is a set of correspondences between entities of ontologies that are semantically related. These correspondences can be used for various tasks, such as ontology integration [1], ontology evolution [2], data integration [3], and data warehouses [4]. Until now, many different tools for matching process are developed. However, most of them suffer from many problems with respect to the quality of matching results. Thus, in this paper we proposed an approach to get a better results from matching processes. The solution is adding a preprocessing phase to matchers. In the preprocessing phase, many input ontologies are analyzed in order to detect inconsistencies and inappropriate patterns modeled by various developers. For detecting these inconsistencies, ontology preprocessing language (OPPL) is used. Then, the refactoring rules are applied on detected patterns to repair the inconsistencies in input ontologies. At the end, assimilated ontologies delivered to matchers for matching process. In this paper for evaluating this work one of the best matcher namely, ASMOV from OAEI is selected. The Ontology Alignment Evaluation Initiative (OAEI) is a coordinated international initiative, which evaluate all matchers every year. As you can see in Figure 1, ASMOV has a good rank in comparison to other matchers with respect to the standard evaluation measurements such as precision, recall, and F-measure [5]. The precision, recall, and F-measure are explained in Section V. The second reason for selecting the ASMOV to evaluate our work is that it can do n:m alignment in contrast to other matchers like, SAMBO, Falcon, DSsim, RiMOM, Anchor-Flood, and AgreementMaker which can do only 1:1 alignment.
which algorithms operate. Algorithms can be classified depending on the data/conceptual models in which ontologies or schemas are described. The matching process can be based on its general properties. In particular, this depends on the approximate or exact nature of its computation. The output of a matching algorithm is related to the form of the alignment. For example, the correspondence between ontology entities is either one-to-one or not. Another dimension concerns is the kind of relations between entities that a system can provide. Most of the matching softwares focus on equivalence (\textit{=}) relation, while a few others are able to provide more expressive results (e.g. subsumption and incompatibility) [7, 8].

C. Alignment
The alignment of ontologies \( o \) and \( o' \) is a set of correspondences between two or more (in the case of multiple matching) ontologies. The alignment is the output of the matching process between the entities of \( o \) and \( o' \). The alignment can be achieved in various cardinalities: \( 1:1 \) (one-to-one), \( 1:m \) (one-to-many), \( n:1 \) (many-to-one) or \( n:m \) (many-to-many).

D. Refactoring
Refactoring is recognized as changes that are made to the internal structure of the software in order to make it easier to understand and to modify without changing its observable behavior.

III. Theoretical Background
To establish suitable semantic correspondences between entities of different ontologies, the integration of the input ontologies is needed. Unfortunately, many ontology matching systems ignore the semantics of the input ontologies in the matching process. Therefore, the matching result is not satisfactory. In this paper, we attempt to combine four apparently distant areas to handle this problem. These areas are: ontology matching, ontology patterns, ontology refactoring, and inconsistency repair. Accordingly, in this section, some research conducted in each of these areas are described.

Research in ontology matching has been burgeoning since the early 2000s. So far, most articles on the ontology matching field have focused on the method of matching processes and have introduced some matchers with diverse approaches. In this section, some matchers which have high ranks in ontology alignment evaluation initiative (OAEI) are introduced. ASMOV (Automated Semantic Matching of Ontologies with Verification) [6] have been applied to the lexical and structural characteristics of two ontologies to calculate the similarity measures. Then the alignment have been verified to ensure that it does not contain semantic inconsistencies. RiMOM [9] is a dynamic multi-strategy ontology alignment framework that combines multiple strategies to improve matching efficiency. The key intuition in this framework is that similarity characteristics between ontologies may vary widely. This approach has considered both the textual and structural characteristics of ontologies. RiMOM is a framework based on risk minimization of the Bayesian
decision systems. It employs multiple ontology alignment strategies and sets a combination weight. Another system is Falcon-AO [10], a practical ontology matching system with good performance that acts based on a number of remarkable features. It is an automatic ontology matching system that uses multiple elementary matchers (V-Doc, GMO and PBM), coordination rules, and the similarity combination strategy. PROMPT [11] algorithm consists of an interactive ontology merging tool and a graph-based mapping called Anchor-PROMPT. Anchor-PROMPT [12] uses linguistic anchors as a starting point and analyses these anchors in terms of the structure of the ontologies. GLUE [13] discovers mappings through multiple learners that analyze the taxonomy and the information within concept instances of ontologies. S-Match [14] is a deductive technique for semantic ontology matching which employs a number of elemental level matchers to express ontologies as logical formulas and then use a propositional satisfiability (SAT) solver to check the validity of these formulas.

Generally, all of the above matching algorithms are classified into two categories: elemental and structural. Elemental level matching techniques compute matching elements by analyzing entities in isolation and ignoring their relations with other entities. Structural level techniques compute matching elements by analyzing how entities appear together in a structure and considering the relation of concepts in taxonomy tree [15].

In recent years, some works on ontology patterns is done [16-19]. Ontology patterns have been used in many fields, but they have rarely been applied in the field of ontology matching. Ontology patterns are mainly inspired by software engineering and knowledge engineering [20]. In the following, some previous works in the field of ontology matching by considering the ontology patterns is described. The paper in [21] involves testing the impact of ontology refactoring on the results of three matcher, namely HMatch, Falcon-AO, and ASMOV. In this paper, some modeling errors via name structure analysis were found and three refactoring operations were applied. By considering semantic structures, authors in [22] analyzed collections of OWL ontologies in order to determine the number of occurrences of several combined name and graph patterns. These structures ranged from simple subsumption to more complex constructions. The goal of this paper is to facilitate automatic alignment among different models by finding such patterns in the given ontologies. In [23], the authors concentrate on detection and mutual matching of semantic structures in ontologies. The authors use the equivalence relation, as well as analyzing homogeneous correspondence. Research in [24] presents a simple method of tracking name patterns over OWL ontology taxonomies. This method helps to detect several probable taxonomic errors and modeling inconsistencies with respect to their set-theoretic interpretations. In [25] authors applied weights to the edges of WordNet hierarchy to improve the semantic word similarity. Furthermore the distance of two words and depth of words in semantic similarity assessment are utilized. This approach can be applied for inconsistencies detection phase of matching process.

Until now ontology refactoring is employed in many different areas [26-29], but the impact of ontology refactoring on the ontology matching field is rarely discussed [21, 23, 30]. In this paper, we focus on this matter. In [26], the authors focus on the detection of anomalies as an important criterion for verification. In this paper, some approaches for the syntactic verification of ontologies are explained and definitions are extended with respect to the existence of rules. Furthermore, novel measures are introduced for detecting the parts of the ontology that may create problems for maintainability. This paper [27] proposed an approach for refactoring multimodal knowledge on the basis of a generic data structure in order to support the representation of multimodal knowledge. Moreover, how this data structure was created from given documents (i.e. the most general mode of knowledge) was explained, along with how different refactoring could be performed by considering various levels of formality. In [29], the authors present the semantic knowledge wiki, Know WE, used to capture and share ontological knowledge for the effective elicitation of problem solving knowledge. Also, a distributed knowledge acquisition process and refactoring phase are shown. In [30], a semi-automatic process for lifting meta-models into ontologies is proposed that allows creating the semantic integration of modeling languages. In so doing, implicit concepts in the meta-model are changed to explicit concepts in the ontology. The application of refactoring patterns on the resulting ontologies could improve automation support for semantic integration tasks. The paper [28] presents a method to develop conceptual schemas as refinements of more general ontologies. For obtaining final conceptual schemas, three activities are performed: refinement, pruning, and refactoring. The refinement phase is done to execute a set of additive operations to the ontology to create necessary elements. Afterwards, in the pruning phase, some unnecessary elements are deleted. Then, a pruned ontology is obtained. At the end, the pruned ontology can be improved by using refactoring operations to obtain the final conceptual schema.

IV. METHODOLOGY

In this part, an approach is proposed to improve the quality of the matching results. The aim of this approach is improving the alignment results by finding the inconsistencies before matching process. Our previous works [31], [32] focus on only lexical and structural patterns, but in this work we concentrate on some new inconsistency patterns. To accomplish this aim, a pre-processing phase is added to matchers. In the pre-processing phase, at first, a comprehensive survey to find the inconsistencies in input ontologies are performed. Then various lexical and structural patterns, which have been modeled by different developers, are detected. Afterward, some refactoring operations are applied on these patterns for repairing
the ontologies. Finally, these repaired ontologies are used as inputs of the matching process. This process is evaluated by ASMOV [33]. Experimental results indicate that better outcomes can be achieved by applying the pre-processing phase as opposed to original ones. In the following the details of this work is elaborated.

A. First step: Inconsistency Detection Phase

In this step, some inconsistencies were detected based on our preliminary analysis of many ontologies. For detecting lexical inconsistencies, the name of entities, especially classes in OWL ontologies are analyzed. The lexical feature consists of all information readable by humans in the ontology. Various ontologies use different methods for defining the names of homogeneous concepts, especially for compound words. In OWL ontologies, different styles in concept naming lead to many obstacles for calculating lexical similarities in matchers.

In ASMOV, three lexical concepts in OWL ontologies are considered: id, label, and comment. ASMOV uses the Lin method [34] for calculating the lexical similarity. As an instance, in two ontologies of a conference track, namely Conference and Ekaw two different class naming for a similar concept is discovered, <Conference#conference-www> ~ <Ekaw#website> and also <Conference#rejected-contribution> ~ <Ekaw#rejected-paper>, both of them couldn’t be found by lexical similarity phase of ASMOV. Therefore, to solve these kinds of problems, some lexical patterns and inconsistencies are detected based on naming ontology design patterns [35] for the purpose of unifying the naming for these different styles of naming. To accomplish this, we used one refactoring operation called renaming operation (RN), which is described in the next section. By doing this, calculating the lexical similarity in matchers, which is done by different methods, can do better than before. Thus, better results can be obtained from the matching process.

Structural patterns are based on the fact that the taxonomic structures of ontologies are often varied and confusing. One reason for this is that different developers have dissimilar viewpoints for developing ontologies. Therefore, they utilize different hierarchies and granularities for defining the entities of ontologies in the same domain. For example, in two ontologies of the conference track namely Conference and Ekaw, realize that there are two different granularities in concept naming for the similar concept _author_. In Conference, three levels of granularity for _author_”-is found which include: contribution_regular-author, contribution_co-author, and Conference__1st-author. However, in Ekaw, there was only one level of granularity for author, namely Paper_author. Furthermore, many problems for calculating the structural similarity by some matchers have been recognized. The relational or hierarchical similarity phase in most matchers is computed by combining the similarities between the parents and children of entities that want to be compared. By considering the problems mentioned above and matcher_work, we realized that different taxonomic structures and different granularities in peer ontologies cause many problems in the matching process. For solving this problem, another refactoring operation, called restructuring operation (RS), is employed for assimilating the structural features of OWL ontologies. Our results show that, in most ontologies, there are significant number of occurrences of the aforementioned patterns.

B. Second step: Refactoring phase

In this phase, by refactoring operations some patterns and inconsistencies, which are detected in previous phase, are repaired. All cases of the modeling errors detected via some patterns mentioned earlier can be repaired by two refactoring operations. The detection of these patterns is the starting point for a refactoring. Generally, refactoring is a process for performing some changes in the internal structure of the software in order to make it easier to understand and to modify without changing its discernible behavior. In this literature, the refactoring process of an ontology matching field is applied. Thus, some changes are done in ontologies by a semi-automatic process. By doing this, new and more understandable versions of ontologies for users and matchers are produced. These versions of ontologies can be utilized more effectively by different ontology matching tools.

There are three general refactoring operations: adding operation (ADD), restructuring operation (RS), and renaming operation (RN). These operations consist of different steps depending on the detected situation [21]. In this paper, RN and RS are used for lexical patterns and structural patterns, respectively. More desirable results in lexical similarity of matchers can be obtained by applying the rename operations for the name of the classes. The rename operations are done by considering the name of the classes in the ontology that have the same taxonomic structures in the peer ontology. Furthermore, by considering the parent-child relations and various granularities used in peer ontologies, restructuring operations are applied for assimilating the structural features of the OWL ontologies. Experimental results show that, better results can be achieved from the structural similarity phase of matchers by transforming a part of ontology into another one. We carry out our experiments on seven pairs of ontologies from the conference track. The reason for choosing these seven pairs among other ontologies is described in the next section.

The number of RN and RS operations applied on these seven pairs of ontologies is explained in the following. In four pairs of ontologies, <Cmt- ConfOf>, <Cmt-Ekaw>, <Conference-Ekaw>, and <Edas-Ekaw>, RN operations are applied more than RS operations, because of the many different lexical patterns find in these pairs. Besides, in other ontology pairs, <Cmt-Sigkdd>, <Conference-ConfOf>, and <ConfOf-Sigkdd>, RS operations are utilized more than RN operations, because these pairs of ontologies have different hierarchical structures and RS operations is used for assimilating the taxonomies.
C. Data set

Some ontologies from OAEI is selected to evaluate this work. The OAEI offers several tracks and subtracks concentrated in different types of matching problems. Our approach was tested on the Conference Track [36]. They are described in OWL-DL and published in the RDF/XML format [37]. This data set is a well-known data set to the organizers and has been used in many ontology matching evaluations. The Conference dataset can be viewed as a much more challenging test cases in contrast to other ontologies of OAEI, such as the Benchmark dataset [38], [39]. Our experiment was carried out on six out of sixteen ontologies of the Conference Track. These ontologies are cmt, confOf, ekaw, conference, edas, and sigkdd. The reason for selecting these six ontologies among others is that reference mapping (also referred to as the gold standard) is available for all possible combinations of these selected ontologies. To evaluate the accuracy of the matching process, it is necessary to determine both the number of correctly found correspondences and the number of incorrectly found correspondences.

D. Implementation

Our implementation is based on the employment of Java language with Jena API in Net Beans IDE. Furthermore, protégé and the Ontology Pre-Processor Language (OPPL) were used for manipulating ontologies written in OWL. OPPL is a domain-specific language, based on the Manchester OWL Syntax. OPPL instructions can add or remove entities and add/remove axioms to entities in OWL ontology. The OPPL Instruction Manager is a Java library that processes OPPL instructions to make changes in OWL ontology. This language is also suitable for defining independent modeling macros that can be applied across ontologies [40].

E. Practical Example

Presented in this section is a practical example to clarify the proposed approach by testing the work with ASMOV matcher. Figure 2 illustrates different styles in class naming and various taxonomic structures for defining the same concepts in a part of two ontologies, namely ConfOf and Sigkdd.

Fig. 2: Different class naming and taxonomic structures using Protégé software [41]

The inconsistencies are located in two peer ontologies are detected with OPPL. After that the ontologies are manipulated by applying the refactoring rules on each one. Then assimilated ontologies are delivered to ASMOV matcher as inputs. The alignment results are shown in figure 3.

Fig. 3: Correspondences found by ASMOV after refactoring [6]

V. Evaluation of the Matching Results

For evaluating the matching results three standard measures, precision, recall, and F-measure is used. Precision is defined as the number of correctly found correspondences divided by the total number of found correspondences. Recall is considered as the number of correctly found correspondences divided by the number of reference alignment. A perfect precision score of 1.0 means that every correspondence computed by the algorithm was correct (correctness), whereas a perfect recall scores of 1.0 means that all correct correspondences were found (completeness).

Precision and recall are defined in (1), (2) [42].

\[
\text{Precision} = \frac{\#\text{Correctly found matches}}{\#\text{Number of all found matches}} \quad (1)
\]

\[
\text{Recall} = \frac{\#\text{Correctly found matches}}{\#\text{Number of reference alignment}} \quad (2)
\]

F-measure represents a trade-off between precision and recall and it is calculated as (3).

\[
F - \text{Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)
\]

Experiments were performed on seven pairs of ontologies from the conference track. The alignments generated automatically by ASMOV for these pairs of ontologies before and after of applying the proposed approach. The results were illustrated in figure 4, figure 5, and figure 6. The results of our experiments show that transformed ontologies improve the matching results with respect to the standard evaluation measures i.e. precision, recall, and F-measure.
VI. CONCLUSIONS AND FUTURE WORKS

In this paper, an approach has been presented for overcoming the uncertainty in the ontology matching results. This work is done by detecting the inconsistencies in ontologies before starting the matching process. In the first step, many problems caused by ontology diversity are resolved. So that, some inconsistencies and unexpected errors, which have been modeled in input ontologies, are detected. After that two refactoring operations, RN and RS, is applied to repair them. This work makes transformed ontologies easier to understand by both humans and matchers. Furthermore, some common mistakes in the alignment results are reduced. The transformed ontologies evaluated with one of the best-ranked matchers, ASMOV. Our experiments were carried out on ontologies of the conference track. Experimental results show that our approach improved the quality of the matching process with respect to standard evaluation measurements, i.e. precision, recall, and F-measure.

For future research, new solutions can be proposed for overcoming the uncertainty and other challenges in the field of ontology matching. Furthermore, our approach can be tested on other matching tools, especially those participating in the OAEI contest. Moreover, some detectable patterns for discovering errors of ontologies and other refactoring operations for repairing them can be extended.

ACKNOWLEDGEMENT

This research was partially supported by Ferdowsi University of Mashhad under the grant number 2/39961 to which the authors likes to extend his thanks to.

REFERENCES


Bahareh Behkamal has received her B.S degree in Computer Engineering from Azad University of Mashhad, Iran in 2005 and her M.S degree in Computer Engineering from Pnu University of Tehran, Iran in 2011. She is currently Ph. D candidate in Computer Engineering at Ferdowsi University of Mashhad, Iran. Her past research interests include Semantic Web and Ontology Matching. She has published some papers in international journals and conferences. She is currently researching in the area on Structural Bioinformatics.

Mahmoud Naghibzadeh has received his M.Sc. and Ph.D. degrees in Computer Science and Computer Engineering, respectively, from University of Southern California (USC), USA. He is now a full professor at the Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran. He is the director of Knowledge Engineering Research Group (KERG) laboratory. His research interests include the scheduling aspects of real-time systems, Grid, Cloud, Multiprocessors, Multicores, and GPGPUs. He is also interested in Bioinformatics computer algorithms, especially protein tertiary structures and protein-protein interactions.
A Framework to Create a Certificate for e-Commerce Secure Transaction Protocol

Nasrin Alishirvani
Department of engineering, Payame Noor university
Tehran, Iran
n_alishirvani@pnu.ac.ir

Received: November 16, 2016- Accepted: March 12, 2017

Abstract—The development of e-commerce requires security to win the confidence of its stakeholders. Among the common protocols to establish safe financial transactions, Secure Electronic Transaction (SET) protocol has more security providing a safe protocol of payments at the level of the communication network between the buyers, sellers, banks and payment gates. In this protocol, all participants in the transaction should receive Certificate Authority (CA) identity. This paper analyzes the secure communication solutions and variety secure communication contracts for financial transactions. Then, architecture is presented to establish a web-based Certificate Authority (CA) identity for elements of Secure Electronic Transaction (SET) protocol and its implementation is described. The presented Certificate Authority (CA) identity in the article can process the requests of entities for online processing and transmit them through Hypertext Transfer Protocol (HTTP).

Keywords: Component, Secure Transaction Protocol; Certificate Authority; Secure Electronic Transaction; Payment Gateway; Public Key Infrastructure

I. INTRODUCTION
E-commerce has reached an important place by introducing information technology in the world and development of the Internet. However, what prevents the rapid growth of the business is consumer perception of security weaknesses on the network. As more payment methods are via credit cards, customers fear from the disclosure of their credit card information.

At the present, Security of financial transactions is supplied by secure communication protocols, especially ones such as SSL / TLS and SET [1]. SSL / TLS protocol is used in wider extent due to its simplicity, while the security of SET protocol is higher than the other one.

SET provides a secure payment protocol in the communication network between the buyer, the seller, the bank, and the payment gateway. In The SET protocol; all participating parties in the transaction must receive a certificate from the CA identity center. The position of this protocol in the protocol stack of TCP / IP is demonstrated in the Fig. 1 [1, 2, and 15].

SET is a set of algorithms and protocols that all are under specific procedure and protocol which enables users to make transactions via a public network such as the Internet. Therefore, to design and implement of SET protocol, identifying and studying of these algorithms,
and how to use them is essential. In this paper, secure communication solutions for financial transactions are investigated, secure communication protocols, including SET, SSL / TLS, IPSec, PGP, etc. are also analyzed.

In the end, architecture for establishing the identity certificate authority for payment gateway, on the basis of SET protocol is proposed.

According to proposed framework, CA Server is in the network for online in order to provide the security in the center from the invalid IP addresses using a three-layer firewall and NAT router and Roll located at the ISP to provide security between different components of SSL protocol.

II. AN OVERVIEW OF THE PAST LITERATURE

So far, lots of works have been done on the design and implementation of PKI. Some of these are available for free on the Internet and some are commercial products [3], including PyCA, OpenCA, Entrust / PKI and IBM Registry. Among them two free cases and two commercial ones are discussed as follows:

- **PyCA**: It is a set of CGI1 scripts that provides the interface between World Wide Web pages (WWW) and identity certificate authority. Scripts have been written in Python. That is why they were named pyCA. pyCA only implements a center of identity certificate and it is not accounted as PKI. Therefore, it does not maintain private keys of individuals and users are responsible for backing up their private key and their certificate [4].

- **OpenCA**: OpenCA is a common effort to create a public key infrastructure. It can be considered as reinventing pyCA in terms of programming but the difference is that its scripts are written in Perl. OpenCA uses OpenSSL as the underlying structure [4, 5].

- **Entrust/PKI**: Family of Entrust company products presents a solution for Public key infrastructure that are needed for e-commerce. This solution includes Cryptography-Based Services and digital signature. Entrust/Commerce CA is a product of Entrust to issue certificates for terminating entity. To do this, it receives certificate from higher certificate authority in the hierarchy of SET [6, 7].

- **IBM Registry**: IBM Registry is part of Net commerce product of IBM company that plays the role of the center of identity certificate in SET protocol. IBM Registry can be used on different machines as distinct identity certificate authorities(CA’s). several CA’s can be also run simultaneously on a machine [8].

III. SECURE TRANSACTION PROTOCOL

With the introduction of information technology in the world and the Internet, E-commerce has reached an important position. One of the principles of e-commerce transactions is secure transaction, secure transactions over the network specially Internet requires high security. As sending data and financial information such as credit card numbers, account numbers, sending confidential financial information, codes, passwords and thousands of confidential pieces of information brings about much concern and this is a justified reason for the significant importance of methods of providing security and different types of secure payment systems[9999].

So far, many protocols have been proposed as secure transaction protocols that only have one or a few features of secure transaction protocols. Some Examples of these are protocols of [10]: PGP, PEM, S/MIME, IPSEC, SSL, STP, and SET are examined.

A. Pretty Good Privacy (PGP)

PGP is the service of confidentiality and authentication for e-mail and file storage applications. Sending Public key is done by identity authentication that its transmission is physically impossible on the network. Sometimes it is electronically transferred or endorsed by telephone [11].

Another method is transferring by a trusted individual who has the public key. For instance public key of user A is signed by recognized and trusted user B and is sent to the user C or the transfer is done by the CA. However public key of individuals are not identified in this way but these CAs are used as accumulator [12].

B. Privacy-Enhanced Mail (PEM)

PEM is one of the algorithms that were created for e-mail security but it was not successful, and it is used rarely to send e-mail nowadays, due to the following reasons:

- PEM uses 56-bit DES encryption which is not a strong cryptography algorithm.

- Also, all the messages were signed in PEM and the signature was situated outside the encrypted message.

---

1 CGI stands for Common Gateway Interface. it is a part of the Web server that provides a feature so that in case of running a program on the server side, its output is displayed for user who is connected to the server through the web page.
and it was not necessary to decode message in order to find signer.

C. Secure Multipurpose Internet Mail (S/MIME)

S / MIME Protocol, is used for secure e-mail. It can be implemented in each automatic transmission mechanism which supports MIME (like HTTP); it uses the security services and requires no user interaction [12].

D. SSL (Secure Socket Layer)

SSL is a solution to communicate between a server and a client securely. The benefit of using this security protocol is taking advantage of its embedded security for securing Non-secure protocols of application layer such as HTTP, LDAP, IMAP, etc. Based on it, the encryption algorithms are applied on raw data which are supposed to pass through Non-secure communication channel as internet and data confidentiality during transmission channel is ensured [1, 13].

In other words, a company that is competent to issue and award of SSL digital certificates, issues special server and client certificates for each of the two sides that are supposed to have secure inter-networks communications, And using its own authentication mechanisms, affirms the identity of each of the parties to the opposite side. Moreover, it should ensure that if information is stolen while being transmitted, it should not be understandable and usable for robber because of using of encryption algorithms and asymmetric and symmetric encryption keys.

E. IPSec (Internet Security Protocol)

IPSec is not the general security architecture for the Internet but a developed protocol on IP protocol which has been provided for IP security. This protocol is widely used in virtual private networks (VPN) in offices related to a company or organization and essentially is implemented in establishing a secure connection between two or more organizations. IPSec uses two protocols of AH and ESP for more confidence on authentication, data integrity and confidentiality. The protocol can establish security at the network layer and in higher layer protocols in Transport state. The main objective of IPSec is augmenting security mechanisms ToIPv4 and IPv6 in order to provide required security for users.

F. SET (Secure Electronic Transaction)

Secure Electronic Transaction (SET) is a security protocol that was developed by Visa and Master Card. Unlike SSL which is an all-purpose system for encrypted communications, SET is only for transactions of credit card between buyer and seller.

SET protocol provides all the facilities supported by SSL. In addition, SET prevents awareness of seller from the buyer's credit card number and only presents it to the bank that issues the credit card. Business needs [14] which are met by the SET [15, 16, and 17] include:

Confidentiality, data integrity, identification and cross-operational capability, in this section, these needs and how they are met by the SET are taken into account [18].

Confidentiality: payment details and how it works should not be visible for someone who secretly monitors this process. Confidentiality is guaranteed by the use of message encryption. Including:

- Symmetric-key cryptography such as: DES, 3DES, RC2, RC5
- Public-key cryptography, like : RSA, Diffie-Hellman, Digital Signature, Message Digest


Data integrity: it should be assured that data received in the SET transaction is exactly the same as sent data. The evolution is guaranteed using digital signatures. To prevent distortion of the message in the route, SET uses digital RSA signature and SHA-1Hash and in some cases HMAC and SHA-1 for ensure data integrity assurance.

Cross- operational capability: SET specifications must ensure that it works in different hardware and software platforms. The ability to exchange information is guaranteed by using the message templates and specific protocols. Content and message encryption is in a standard form. SET utilizes ASN.1 and DER for this purpose. It uses the standard of PKCS #7 to encapsulate the messages.

IV. THE PROPOSED CERTIFICATE AUTHORITY (CA)

A. Tasks of the suggested CA identity

The tasks of CA identity in the SET protocol include: issuing certifications, canceling certificates, and maintenance of certificates, maintenance of lists of denial of certifications, and Brand CRL2 Identifier (BCI). Since in the protocol, private key entities in the end are not sent to the CA identity, it does not take responsibility for private key entities in the end. And each entity is only responsible for maintenance of its own private key (s).

In SET protocol, tasks of CA identity center of the card holder are similar to the CA identity centers of seller

\[2\] Certificate Revocation List
and payment gateway. Both CA identity centers of card holder and seller are responsible for issuing certifications, canceling certificates and maintenance of certificates, maintenance lists of denial of certifications, and BCI.

- In the proposed model, the CA identity center of payment gateway takes responsibility for issuing and keeping certificates, lists of denial of certifications and BCI. We also have certified payment gateway falsified, unlike the other end entities.

B. Features of the identity certificate authority of proposed payment gateway

In this architecture both valid and invalid IPs are used to enhance the security. Unlike the usual way, valid IP in local network of CA server is not set on any computer and invalid IP range within one class is used in all networked computers and other devices on the network, and also the router which is connected to the ISP suffering and wishes of a class.

To receive sent packets to this network with a valid address, a rule is used which has been programmed in the router located at the service-provider. Therefore, packets with the destination address (Valid Address) and Address of origin (Any Address) are sent to the local network. A firewall with three layers at the local network gateway is situated with the possibility NAT.

Activities which are done at any layer of the firewall are as the following:

- **Access Control**
  The first layer firewall: To determine the prohibited packages (black) and delete them. Header fields of IP packet are analyzed in the first layer.

  Source address: some of the machines inside or outside the network do not have the right to send the packet, therefore their packet are deleted upon arrival at the firewall.

  Destination address: some of the machines inside or outside the network do not have the right to receive the packet, therefore their packet are deleted upon arrival at the firewall. Invalid IP address in transmission mode in case of sending packet to the local network is any IP except valid IP of local area network.

  The second layer of firewall: closing some ports of some services, such as Telnet, FTP, etc.

  In the second layer, header fields of transport layer are examined. Due to the fact that Port number of the process of source and destination are the known standard port numbers, only packets with ports numbers that CA Server will listen to them are passed and the rest of the packets are deleted. However other ports can be closed except intended ports of CA Server.

- **Filtering**
  The third layer of firewall:
  Analyzing the outcome of a web page text, E-mail, and so on, due to the fact that protection in the third layer is made based on type of service and the application, packets are controlled by controlling standard protocol of X.509.
  For example, a request that its Issuer Name field is not same as the name of our CA or other acceptable CA is thrown away. Therefore in this level, processing volume is high.
  - Network Address Translation (NAT)
    After the operation of Access Control and Filtering, valid address of CA Server must be translated to invalid address in the internal network. Thus the packets passed the NAT and the reverse operation is done for traffic outside the internal network.
  - To enhance the security in this architecture, different parts of the local network are associated together via native SSL protocol; manipulated SSL.
  - In the proposed center of certificate for communicating between users and identity certificate authority, HTTP protocol in secure form of HTTPS is used.
  - Ability to request certificate online exists in offered identity certificate authority where CA is maintained on-line and is capable of processing requests online.
  - LDAP directories are implemented to publish and retrieve certificates and revocation lists of certificate. This directory enables a secure connection and information is securely exchanged between this directory and the CA.
  - The identity certificate authority of Payment Gateway is responsible for issuing and maintenance of certificates and certificate revocation lists and BCI . moreover, against other terminating entities, Certificate Revocation of Payment Gateway is also available.

C. Architecture of proposed authentication center

Considering the duties of identity authentication centers of card holder, seller and payment gateways of identity authentication centers is composed from the following components Fig. 2:

- **CA Server**: It plays the role of identity Certificate Authority. All requests for issuing certificate is given to this server. If CA Server requires identification of any request, sends a message to the RA Server.

- **RA Server**: It is responsible for identification of individuals. When RA Server receives request of entity identification, identifies that one using the information in its database identification And announces the result is to the CA Server.
LDAP Repository: LDAP directory is used for maintaining certificates and certificate revocation lists. CA Server certificate places certificates and received certificate revocation lists from higher CAs in the hierarchy of SET as well as certificates and certificate revocation lists issued by itself in LDAP directory.

Fig. 2: Architecture of the Proposed Identity Certificate Authority

D. The Process of Receiving Certificate by Users in the Proposed Architecture

The certificate process certificate for seller and receiver is very similar to this process for the cards holders, but there are significant differences that arising from the nature of the parties involved in the transaction that include:

- The seller has two pairs of keys (A pair for signs and another pair for key exchanges).
- The certificate process for card holders will probably perform on the Internet and on-line entirely, while in the case of seller there is always offline validation too.

In this paper the certification process for card holder is explained in details:

The card holder begins the registration.

The card holder is connected to the RA web server through its browser and requests for awakening message. By receiving the awakening message and informing of the URLs needed to connect to the CA, the certificate request process will begin. The card holder sends the beginning message through HTTPS protocol to the RA server.

RA server sends the response.

RA server receives the beginning request of the card holder and after investigation sends it to the CA server. CA server also sends both itself certificates for the public signature key and key exchange of public key. These certificates will be used to protect payment card account number on the request for registration form.

The card holder requests the registration form.

Identity authentication center needs some details to confirm that cardholder is the same person who claimed. The card holder receives the beginning response message from RA server and measures the accuracy of the certificates by examining the chain of trust to the roots. The accuracy of identity authentication center’s sign measured by performing the beginning response (as shown in fig. 3 through a hash function and a summary message is created. Digital signature is decrypted by using public signature key of Identity authentication center and its results are compared locally to the obtained abstract message. If they are the same, the accuracy of Identity authentication center’s response is guaranteed.

The card holder enters the credit card account number and the card holder's software creates a registration form. Request a registration form is encrypted by the symmetric key (K1 in Figure 3) that is randomly created. These symmetric key and card holder’s account number are encrypted by key exchange of public key and a digital package is created.

Identity authentication center sends registration form

Registration form is transmitted emptily, but Identity authentication center signs the message in order to the card holder be able to ensure of its validity. RA server receives the registration form, since previously identification has been conducted sends the digital package to the CA server. CA server decrypts the digital package by key exchange of private key of identity authentication center in order to achieve card holder account number and symmetric key. Symmetric key is used to achieve the request of the registration form (via decryption of encrypted registration form). Identity authentication center determines the appropriate
registration form for this user and a digital signature is created (as shown in Fig. 3 - a). Identity authentication center sends the registration form, digital signature and certificate of Identity authentication center (containing the public signature key) to the card holder.

Card holder requests the certification
The card holder’s software provides the request form and creates a request message. The card holder receives message of registration form from the RA server and measures the accuracy of certificates by examining the chain of trust to the root and based on fig. 3 the accuracy of message are guaranteed. The card holder’s software creates a pair of keys and stores them safely then fills out the registration form and sent it along with a random number in order to create a certification by identity authentication center. It creates a certificate request message by using the information in filled registration form and random number. Certificate request and registration form are encrypted by private signature key of card holder (as shown in Fig. 3) and a digital signature is created. The card holder creates two keys accidentally (we call them symmetric key (1) and symmetric key (2)). The card holder catches registration form, certificate request, digital signature, public signature key of card holder and the symmetric key (1) and encrypts them by using symmetric key (2). Symmetric key (2), the card holder’s software information, history of accurateness and random number encrypted by key exchange of public key of identity authentication center and a digital package is created. Digital package and encrypted message are sent to the RA server.

Identity authentication center receives the request
Now Identity authentication center should decrypt the message sent by the card holder and process the registration request. Digital package is decrypted by using key exchange of private key of identity authentication center in order to reveal Symmetric key (2), account information and random number. Encrypted message is decrypted by using a symmetric key (2) in order to the public signature key card holder, the symmetric key (1), certificate request, registration form, and digital signature be achieved. The authenticity of the signature of the card holder is guaranteed by performing certificate request and registration form (as shown in Fig. 3) and by using public signature key of card holders. Now Identity authentication center reviews the certificate request by verification of card holder’s account and details of registration form in comparison with information of trademark database. If the information is valid, RA server sends its request to the CA server. In CA server the following steps are passed.

First, identity authentication center creates a random number that combined with the random number that created by the card holder’s software in order to create a serial number. Account number, expiration date and serial number are encrypted by using one-way hash algorithm. The obtained result is inserted in card holder’s certification. Then the certification is signed digitally by identity authentication center and by using the private signature key. The certification placed in a certification response message which is signed by passing it through a hash function. The abstract of message that created in this way is encrypted by private signing key of identity authentication center that it will bring a digital signature. Identity authentication center encrypts the digital signature and response of certification that contains a random number (that created by identity authentication center) and other information (such as logo trademark), with a symmetric key (1) that the card holder receives. The obtained message, certification of identity authentication center and the certification of card holder are sent to the card holder.

Card holders will receive a certificate
The card holder reviews the accuracy of received certification and stores it. The card holder reviews the accuracy of certification by examining the chain of trust to the root. The encrypted certification response is decrypted by using symmetric key (1) in order to obtain the certification response and digital signature. The card holder’s software now recognizes the symmetric key (1) since created it in first place. It combines the random number sent in the registration form with the random number in the certification response in order to obtain the serial number. This number is stored for use with a certificate. Accuracy of received signature of identity authentication center is guaranteed by performing the certification response (Fig. 3) and by using public signing key of identity authentication center. The card holder’s software stores the certification in a safe place on disk or another intermediary in order to use it in the future. The card holder can buy now. This process is shown in Figure 4.

V. Evaluation of performance
Parameters considered in the assessment of identity certificate authority include: Throughput and average response time for requests of opening, registration forms and certificate. Throughput is referred to the number of requests processed per unit of time. The Throughput chart of identity CA based on the number of concurrent Applicants is shown in Fig. 5. throughput of the center is slightly reduced by increasing number of applicants.

In contrast to other servers, CPU time is wasted here and efficiency is being reduced by declining the number of clients while the performance is being boosted by increasing of demands.

In this server, such state does not exist because the efficiency is not augmented by increasing of clients due to cryptographic operations and filling of CPU time. Average response time of requests is referred to the average time servers take in order to respond to each question. Fig. 6 demonstrates Average response time of certificate request form.
VI. THE COMPARISON OF THE PROPOSED ARCHITECTURE WITH PRESENTED ARCHITECTURES

PyCA architecture uses a pair of keys, while the SET protocol requires two pairs. In this architecture, the private system of CA is kept is offline. Lots of messages to users in the SET protocol are identity signed by the certificate authority. This would be in conflict with the lack of availability of CA private system. In this architecture, much work has not been done for the identity of users. CA server in OpenCA architecture is kept as offline and various stages of certifying require user interaction, which is accounted as an architectural weakness. This architecture places a strong emphasis on the identity of individuals. OpenCA is capable of publishing certificates and Certificate Revocation Lists in LDAP directory. Usability of the issued certificate by other identity certification authorities exists in the architecture. Therefore, it can be implemented for hierarchical model. The architecture of the certificate revocation list is used to declare the status of individuals. Architecture of Entrust / PKI does not have previously expressed weaknesses. Entrust / Commerce CA product of this company is written specifically for SET protocol. Entrust / Commerce CA just supports some identity certificate authorities of the SET protocol. IBM Registry is part of the Net Commerce product of IBM Company and purely has been written for SET protocol. This production also supports only some identity centers in The SET protocol. As the features of the proposed identity center mentioned in the article, Comparison between the proposed identity CA and existing identity centers (commercial and noncommercial ones) in terms of capabilities has been conducted in Fig 7.
VII. CONCLUSION

In this paper, secure transaction protocols and their features were investigated. A number of commercial and non-commercial identity certificate authorities were discussed and according to the SET protocol requirements of an identity certificate center devoted specifically for this Protocol, an identity certificate authority was presented. The proposed identity certificate authority is capable of processing entities Requests online. And requests are sent and received through the HTTP protocol. As the CA server is online on the network, invalid IP addresses, three-layer firewall, NAT and router rule located at the ISP are used to establish security in the identity certificate authority. SSL protocol has been implemented between different components to provide security.

References


[3] Booz & Allen, Public Key Infrastructure (PKI) X.509 Certificate and CRL Extensions Profile, Elkridge Landing Road, Annapolis, Maryland, October 12, 2005


Nasrin Alishirvani was born in Tehran, Iran, in 1974. she is currently faculty member at the Department of engineering, Payame Noor university. She authored the Computer networks 2 book. She has published a number of research papers in international scientific journals and conference proceedings.
GSSC: Graph Summarization based on both Structure and Concepts

Nosrat Ali Ashrafi Payaman
Computer Engineering
Iran University of Technology and Science
Tehran, Iran
ali_ashrafi@iust.ac.ir

Mohammadreza Kangavari
Computer Engineering
Iran University of Technology and Science
Tehran, Iran
kangavari@iust.ac.ir

Received: January 10, 2017- Accepted: March 26, 2017

Abstract— In this paper, we propose a new method for graph summarization named GSSC, Graph Summarization based on both Structure and Concepts. In this method, an attributed graph is summarized by considering both of its topology and related concepts. In this method, for a given attributed graph a new graph is constructed that an edge in this new graph represents structural and conceptual similarity of its two end points. Structural and conceptual similarity of two nodes not necessarily has the equal amount of importance in the weight of the resulting edge. For example, for a special case such as query answering, structure can be more important and vice versa. Similarity of two nodes is computed based on Jaccard similarity. This method has some advantages such as flexibility, simplicity, learning capability, user-orientation that makes it a better method for graph summarization. We implemented our method and the method proposed by Bei and evaluated these two methods on real-life dataset HEP_TH. Our experimental results showed effectiveness and efficiency of our proposed method.

Keywords- Graph summarization; super-node; similarity; conceptual summarization; summary.

I. INTRODUCTION

Graphs are used in a variety of applications for modelling data and their relationships. Examples of data modelled by graphs include social networks, communication networks, web graphs, biological networks, chemical compounds, etc. Graph theory and its applications has attracted the attention of the scientists [1] and specially there is a survey of existing work on graph matching, describing variations among problems, general and specific solution approaches, evaluation techniques, and directions for further research [2]. In that survey an emphasis is given to techniques that apply to general graphs with semantic characteristics.

These days many applications generate large scale and massive graphs with billions of nodes and edges and a lot of research has been done on theory and engineering of Tera-scale graphs [3]. In fact, we are faced with graphs which are very massive and their growth rate is also increasing rapidly. For example Facebook has had 1.11 billion members on March 2013 while at the end of 2004 had only about 1 million members (http://news.yahoo.com/number-active-users-facebook-over-230449748.html).

Query answering on these massive graphs is very time-consuming. Graph summarization has proposed as a solution for this problem. Recently several graph summarization algorithms [4], [5], [6], [7] have been proposed that reduce a massive graph to a smaller one by removing its details but preserving its overall properties. This smaller graph can then be used for query answering. Of-course these answers are not exact and have a little error. This error is acceptable because of lower response time which is a necessity in a lot of applications.
The formal definition of a summary graph according to [7] is as follows:

Definition 1. (Summary Graph) Let \( G = (V, E) \) and \( \Phi = \{V_1, V_2, V_3, ..., V_k\} \) is a partition of \( G \) such that \( \bigcup_{i=1}^{k} V_i = V \) and \( \forall i \neq j: V_i \cap V_j = \emptyset \). The summary of \( G \) based on \( \Phi \) is \( S = (V_s, E_s) \) where \( V_s = \Phi \) and \( E_s = \{(V_i, V_j) | \exists u \in V_i \wedge v \in V_j \wedge (u, v) \in E\} \).

Fig 1 shows an undirected graph with 8 nodes and 13 edges and one of its summaries. In fact, a graph is partitioned into some parts each containing some nodes of the original graph. Each part is called a super-node. For example, as is shown in the Fig 1.a, the vertices \( a \) and \( b \) in the original graph are grouped together and make a super-node (blue one) in the summary graph (Fig 1.b). The edges between 2 super-nodes are also grouped together and shown by an edge in the summarized graph called a super-edge. For example six edges \( (a,d), (a,e), (b,d), (b,e), (c,d) \) and \( (c,e) \) are packed and shown by a super-edge between two blue and red super-nodes in the summary.

The summary graph has four super-nodes according to four dashed ovals in the original graphs as shown in Fig 1.b. To illustrate more, the super-nodes have the same color as their corresponding groups in the original graph.

A super-edge in the summarized graph shows that an edge must exist in the original graph between a node of the first super-node and another node from the second super-node. For example the super-edge between the red and blue super-nodes shows at least one of the edges \( (d,f), (d,f), (e,f), (e,f), (c,e) \) must exist in the original graph. The super-edge between the red and yellow super-nodes indicates that one of the edges \( (d,f) \) and \( (e,f) \) must exist in the original graph. Here the edge \( (e,f) \) exists in the original graph.

The formal definition of a good summary according to [7] is as follows:

Definition 2. (Good Summary) Let set of nodes in super-node \( V_i \) participate in the relationship \( (V_i, V_j) \) as \( P_{i,j} = \{u \in V_i \wedge \exists v \in V_j \ such \ that \ (u, v) \in E\} \). The participation ratio of the relationship \( (V_i, V_j) \) is defined as \( r_{i,j} = |P_{i,j}| / (|V_i| + |V_j|) \).

For a group relationship, if its participation ratio is greater than 50\%, then we called it a strong group relationship; otherwise, we call it a weak group relationship. Note that in an ideal summary, the participation ratios are either 100\% or 0\%. The participation ratios of a good summary are near to either 100\% or 0\%.

In fact a good summary of a big graph is the one that can be stored in the memory and the more important it should generate answers to the queries the same as answers generated from the original graph. This is the main challenge of the graph summarization.

The majority of summarization algorithms generate structural summaries while most real world graphs are attributed graphs in which every node or edge has a lot of attributes. In this kind of graphs, node attributes are important and must be considered in summarization.

Usually users are interested in summarization based on concepts (attributes). For a given attributed graph, a lot of summaries can be produced according to selected attributes.

The formal definition of an attributed graph is as follows:

Definition 3. (Attributed Graph) An attributed graph is defined as 4-tuple \( G = (V, E, \Sigma, F) \) where \( V = \{v_1, v_2, ..., v_n\} \) is a set of \( n \) nodes, \( E = \{(v_i, v_j) | 1 \leq i, j \leq n \ and \ i \neq j\} \) is a set of \( m \) edges, \( \Sigma = \{a_1, a_2, ..., a_t\} \) is a set of \( L \) attributes. Attributes of a node \( v_j \in V \) is denoted as \( [a_1(v_j), a_2(v_j), ..., a_t(v_j)] \) where \( a_i(v_j) \) is an observation value of \( v_j \) on attribute \( a_i \). The set \( F = \{f_1, f_2, ..., f_t\} \) denotes a set of \( L \) functions and each \( f_i : V \rightarrow \text{dom}(a_i) \) assigns each node \( v_j \in V \) an attribute value in the domain \( \text{dom}(a_i) \) of the attribute \( a_i \) (\( 1 \leq i \leq L \)).

The formal definition of hybrid summarization (summarization based on both structure and attribute similarities) is as follows:

Definition 4. (Hybrid Summary) For a given graph \( G = (V, E) \) let:

1) Every node has attribute set \( A_v = \{a_1, a_2, ..., a_t\} \).
2) \( \Phi = \{V_1, V_2, ..., V_k\} \) is a partition on \( V \).
3) User is interested in attributes \( A_u = \{a_{i_1}, a_{i_2}, ..., a_{i_p}\} \) where \( A_u \subseteq A_v \).

Then a hybrid summarization is \( G_s = (V_s, E_s) \) where:

1) \( G_s \) is a structural summary as previous.
2) All vertices inside \( V_i \) have the same value for all attributes in \( A_u \).
3) The density of edges inside each \( V_i \) is more than a given threshold.
4) The edge density of edges between super-nodes \( V_i \) and \( V_j \) is less than a given threshold.

For some applications, conceptual summarization is necessary for analysing massive graphs, but in general a summary resulted from both structure and concepts may be useful. We propose a method for this purpose.

There is a method [8] for graph summarization based on both structure and concepts. This method unlike our method, at first summarizes a graph based on concepts or similarity of nodes and then tries to adjust the summary with the graph structure. In fact contribution of graph structure to summary is added to the summary after construction of conceptual summary. Zhou et al. [9] proposed an algorithm for graph clustering based on both structure and node similarity. In this method unlike our method similarity of two nodes is measured based on number of random walks between two nodes.

Henceforth our target is to have a method for graph summarization that generates structural or conceptual summaries or even a mixture of these two summaries.
with the arbitrary degrees of contributions. Such a method can be very useful in query answering based on learning degree of contributions of structure and attribute similarities in resulting summary. The more summary is realistic the more queries can be answered precisely. By learning the value of $\alpha$ (structure contribution in summary) based on trained dataset it can produce a realistic summary.

**Motivating Applications:** Graph summarization problem can be motivated by revealing biological modules [10], provenance systems [11] and many other applications. In the following, we further discuss these two applications.

Finding biologically meaningful modules in a network of proteins is important. In fact detection of protein complexes and prediction of biological processes can discover the global organization of the cell. Graph summarization can be used for this purpose.

Provenance systems produce provenance graphs that can be used for tasks such as determining the inputs to a particular process for debugging entire workflows executions. Visualization can be used to support such tasks. By summarization, it is possible to visualize such massive graphs.

The rest of the paper is organized as follows. Related works is reviewed in Section 2. In Section 3 the new proposed method for graph summarization is presented. Evaluation and Experimental results are given in Section 4. Section 5 is dedicated to discussion and finally we concluded this paper in Section 6.

II. RELATED WORKS

In the following, some of recently proposed summarization algorithms are described shortly to illustrate the scope of the problem.

Navlakha et al. [5] proposed a summarization algorithm in 2008 where graph compression is done by collapsing a set of similar nodes into super-nodes and defining a super-edge between every pair of super-nodes. It tries to construct a compression graph with the minimum representation cost based on the MDL\(^1\) idea.

For this purpose, they developed two iterative algorithms, GREEDY and RANDOMIZED. The GREEDY algorithm selects, in each stage, the best pair of nodes to merge based on the representation cost reduction. It is obvious that the running time of this algorithm is high. To reduce the running time, a RANDOMIZED algorithm has been proposed by the authors (Navlakha et al.). Unlike the GREEDY, in this algorithm two merging nodes are selected randomly.

In 2008, Tian et al. [6] proposed a summarization method with two summarization operations called SNAP2 and k-SNAP for grouping nodes and constructing summary. This summarization algorithm has been proposed for attributed graphs. Tian et al. defined attribute compatible grouping and also relation compatible grouping. In addition, they improved the SNAP operation by proposing k-SNAP, where $k$ is the right size of resulting summary and is given by the user.

In 2009, Zhang et al. [7] have improved the k-SNAP operation in two ways. In fact, k-SNAP method has two shortcomings. First, users have to categorize the attribute values and second there is no criterion to measure the quality of the resulting summary. For these shortcomings, Zhang has proposed the CANAL algorithm to categorize attribute values automatically and a criterion to estimate the quality of the summary.

In 2008, Chen et al. [12] proposed the OLAP framework which provides OLAP like operations on graphs. The OLAP framework has been introduced to create cubes from graphs based on dimensions and measures. The natural property of OLAP framework is that constructs a summary based on the selected attributes and given input information.

Another summarization method has been proposed by Chen et al. [13] in 2009 for mining frequent patterns. This method works by producing randomized summary graphs. In fact, Chen et al. confirmed, in the case of massive graphs, that the traditional pattern mining algorithms are very time-consuming and inefficient because of random access time. Therefore, they proposed a summarization method that first constructs summaries and then mines them instead of mining original disk-resident graphs.

In [14] a method has been proposed for graph summarization that guarantees the quality of the summary. This method produces a summary that

\(^1\) Minimum Description Length
minimizes the reconstruction error. The error is computed based on the difference between the adjacency matrices of the original and summary graphs. The authors have presented a connection between graph summarization and geometric clustering. Based on this connection, they have developed a polynomial-time algorithm to compute the best possible summary with a given size.

Navlakha's algorithm which was described as the first algorithm in this section is a bottom-up algorithm. The top-down approach starts with a minimum number of super-nodes and then iteratively splits super-nodes to achieve a summary with a right number of super-nodes. In this approach, a criterion is necessary to divide a super-node. In [6] there is an algorithm (Algorithm 2) to summarize graph based on top-down approach. These days the majority of real applied graphs are attributed graphs such as social networks and web graphs. Recently a lot of papers have been published on attributed graphs. For example we can list papers for matching patterns [15], matching graphs with fuzzy attributes [16] and predicting links and inferring attributes on a social attribute network [17].
Therefore at present, hybrid summarization is more important than structural summarization. All above mentioned summarization methods are single-process solutions and as a result cannot scale to large graphs. In [18] three distributed graph summarization algorithms have been proposed.

Dynamic graphs and their interpretations are also important and in [19] these graphs have been studied, formulated and a new method is proposed for finding coherent and temporal patterns in these graphs.

Summarizing a graph based on both structure and attributes is important and a method [8] has been proposed for this purpose. The method summarizes a graph by introducing real and virtual links. In fact for a given graph, a new graph is constructed where the weight of an edge in this new graph is resulted from both real and virtual links.

III. PROPOSED SUMMARIZATION METHOD

The new proposed summarization method covers two above mentioned kinds of summarization, structural and attribute-based. In fact both structure and concepts (attributes) of the graph have contributions in making the resulting summary. For this reason, we consider two kinds of edges namely structural and conceptual edges. A structural edge is as previous and indicates that two vertices are connected, while a conceptual edge shows the similarity of two vertices based on their attribute values. In this new proposed method, for a given graph a new graph with the same vertices but with new edges is constructed. In the constructed graph, edges are weighted and some edges may be added because of attribute similarities of vertices. Weight of an edge is summation of structural and conceptual weights, of course weighting factors may be different. For more demonstration, let consider an attributed graph in Fig 2.a with similarity of vertices as given in Table 1.

<table>
<thead>
<tr>
<th>Source</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Destination</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Similarity</td>
<td>0.9</td>
<td>0.8</td>
<td>0.2</td>
<td>0.0</td>
<td>0.5</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Based on the given graph, two new graphs constructed. In the first new constructed graph, structural and conceptual similarity has equal contributions in the weight of an edge (\(1\) and \(2\)). In the second constructed graph, the contribution of conceptual similarity is two times as structural similarity (\(\frac{2}{3}\) versus \(\frac{3}{3}\)).

For more illustration, let consider another example as shown in Fig 3.a. Graph in Fig 3.a is the original graph and its new constructed graph is shown in Fig 3.b. In the new graph, the edge \((v_2, v_3)\) has added because of similarity of \(v_2\) and \(v_3\), similarity of these two nodes based on their attributes. Both of them study physics. In this case, contributions of structural and conceptual weights have considered equal.

Structural and conceptual summaries with two super-nodes for student graph shown in Fig 4.a and 4.b respectively. Based on the new proposed method, student graph can be summarized as shown in Fig 4.c.
Jaccard similarity measure is used. Compared exactly while for multi-valued attributes, calculation differs. The single-valued attributes are depending on being single or multi-valued attribute, its given attribute, we use the Formula 4. In fact, weight of an edge is calculated as follows:

\[
wx(e) = αwx(e) + (1 - α)wx(e)
\]

where \(0 ≤ α ≤ 1\)

Where \(α\) and \((1 - α)\) are weighting of structural and conceptual similarity and show their contributions in the weight of an edge in the new constructed graph.

### B. Proposed Method Computations

Weight of edges has an important role in summarization and for this reason the way these weights are calculated must be explained clearly. Weight of an edge is calculated as follows:

\[
w(e) = α \cdot w_{\text{si}}(e) + (1 - α) \cdot w_{\text{st}}(e)
\]

where \(\alpha\) is calculated as follows:

\[
w_{\text{si}}(e_{ij}) = \sum_{l=1}^{k} c_l w_{\text{si}}(e_{ij}, l)
\]

In fact, \(w_{\text{si}}(e_{ij})\) is the similarity of two vertices \(v_i\) and \(v_j\) based on the given attribute set \(\{a_1, a_2, ..., a_k\}\) with the importance degrees \(\{c_1, c_2, ..., c_k\}\) is calculated as follows:

\[
w_{\text{st}}(e_{ij}) = \begin{cases} 0 & \text{if } W[i][j] = 0 \\ 1 & \text{if } W[i][j] = 1 \end{cases}
\]

Where \(W\) is the adjacency matrix of the input graph.

Attribute-based similarity of two vertices is also a reason for overall similarity of two vertices. Attributes can be single or multi-valued. Similarity of two vertices based on the given attribute set \(\{a_1, a_2, ..., a_k\}\) with the importance degrees \(\{c_1, c_2, ..., c_k\}\) is calculated as follows:

\[
wx(e) = \begin{cases} 0 & \text{if } a_i \text{ is single_valued} \land \text{val}(v_i, k) = \text{val}(v_j, k) \\ 1 & \text{if } a_i \text{ is multi_valued} \land \text{val}(v_i, k) = \text{val}(v_j, k) \\ \text{val}(v_i, k) \cap \text{val}(v_j, k) & \text{if } a_i \text{ is single_valued} \land \text{val}(v_i, k) \neq \text{val}(v_j, k) \\ \text{val}(v_i, k) \cup \text{val}(v_j, k) & \text{if } a_i \text{ is multi_valued} \land \text{val}(v_i, k) \neq \text{val}(v_j, k) \end{cases}
\]

For calculating similarity of two vertices based on a given attribute, we use the Formula 4. In fact, depending on being single or multi-valued attribute, its calculation differs. The single-valued attributes are compared exactly while for multi-valued attributes, Jaccard similarity measure is used.
C. GSSC Algorithm

After constructing the new weighted graph, the graph can be summarized in top-down approach. In every step, edges with weight less than a given threshold are removed and graph is partitioned into some subgraphs. This trend continues to achieve a summary with the right size. We have presented this approach in algorithm 1.

D. Super-edge Weight Computation

The weight of a super-edge is computed based on weight of edges between nodes in two super-nodes. For two super-nodes with m and n nodes the weight of super-edge between these two super-nodes can be computed as follows:

\[ w(se_{i,j}) = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} w(e_{i,j}) \]  \hspace{1cm} (5)

Where \( w(e_{i,j}) \) is the weight of the edge between two super-nodes \( S_i \) and \( S_j \). Edge weight is summation of structural and conceptual weights which have contributions \( \alpha \) and \( (1-\alpha) \) respectively. There are some distance/similarity measures such as cosine, n-norm, Jaccard, etc. that can be used for this purpose. Based on application and the aim of summarization, one of these similarity measures can be used. In some situations these similarity measures can be customized. The output of comparing two nodes is a number in interval \([0 \ldots 1]\).

E. TIME COMPLEXITY

In the proposed method at first the weight of edge between every two vertices is calculated and after that the summary is generated by removing edges from less weighted toward high weighted. Thus the time complexity of the proposed algorithm is \( O(|E| \times |V| \times |V'|) \). This time complexity is for the worst case and in the best case when a large number of edges are removed in each iteration of algorithm, the time complexity is \( O(|V| \times |V'|) \).

Algorithm 1 Summarization (G, k, A, \( \alpha \), C)

**Input:** G: graph, k: the right size of the summary, A: user interested attributes, \( \alpha \): the contribution of structure in the resulting summary, C: importance degrees of attributes

**Output:** S: the resulting summary

1. Calculate the weight of edge between every two vertices as \( E' \).
2. Construct a weighted graph \( G' = (V, E') \) where \( E' = \{(v_i, v_j) \mid v_i \text{ and } v_j \in V \} \) based on \( G = (V, E) \).
3. initialize \( w_1, \text{ size and } \Delta w \);
4. \( \text{num} = |\{G' \mid \text{where } G' \text{ is a connected component of } G \text{ and } |G'| > \text{size}\}| \)
5. while (\( \text{num} < k \)) {
6. \( E' = E' - \{e_{ij} \mid w(e_{ij}) < w_t \} \)
7. \( w_t = w_t + \Delta w \);
8. recalculate \( \text{num}, w_{\text{deg}} \) and \( w_{\text{path}} \)
9. }
10. Select the \( k \) biggest connected components as super-nodes and make the summary graph.

IV. EVALUATION AND EXPERIMENTAL RESULTS

To evaluate the proposed method, we selected a real-life HEP-TH dataset and implemented our method and SGVR method and run them on a system with configurations given in Table 6. The details of dataset, application and system are demonstrated in following subsections.

We compared our proposed method with the recently published paper on this subject which is Bei’s method (SGVR). Authors of SGVR method have compared their method with other methods only based on density. Therefore for a fair comparison, we compared our proposed method with Bei’s method based on density. Of course for comparing methods that summarize a graph based on both structure and concepts it is reasonable to compare summaries by considering both density and entropy.

A. Dataset

We considered real-life dataset HEP-TH, which presents information on papers in high-energy physics, for evaluation our proposed method. This dataset is an attributed graph which can be downloaded from knowledge discovery laboratory. Every vertex of this graph is one entity of type paper, journal, author or email_domain. The vertices are connected by attributed edges. The more information about this graph is given in the Table 3.

The number of each entity given in Table 4 and the number edges between every pair of these four entities given in Table 5.

Value of \#\( e_{x,y} \) shows the number of edges between two entities representing by x and y. In fact sub-scripts x and y represent two first letters of four entities paper, author, journal and email_domain. For example pa, au, jo and em show paper, author, journal and email_domain respectively.

For summarization purposes, we considered a subgraph of this graph that only contains nodes of paper type. This subgraph has 29555 nodes.
Table 3: The dataset information

<table>
<thead>
<tr>
<th>#vertices</th>
<th>#edges</th>
<th>#vertex attributes</th>
<th># edge attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>42319</td>
<td>532430</td>
<td>39</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 4: Number of entities in the dataset

<table>
<thead>
<tr>
<th>#paper</th>
<th>#author</th>
<th>#journal</th>
<th>#email_domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>29555</td>
<td>9200</td>
<td>448</td>
<td>3116</td>
</tr>
</tbody>
</table>

Table 5: Number of edges between every pair of entities

<table>
<thead>
<tr>
<th>#e_{pa,au}</th>
<th>#e_{pa,jo}</th>
<th>#e_{pa,Em}</th>
<th>#e_{au,jo}</th>
<th>#e_{au,Em}</th>
<th>#e_{pa,pa}</th>
<th>#e_{au,au}</th>
<th>#e_{jo,jo}</th>
<th>#e_{Em,Em}</th>
</tr>
</thead>
<tbody>
<tr>
<td>58515</td>
<td>20826</td>
<td>0</td>
<td>0</td>
<td>12487</td>
<td>0</td>
<td>352807</td>
<td>87794</td>
<td>0</td>
</tr>
</tbody>
</table>

For each paper, we considered three attributes: num_revisions, downloads_60days, and area. The number of nodes that has each of these attributes are 29555, 1566, and 3199 respectively. Two first ones are numerical attributes and the third one is an alphabetical field. Conceptual similarity of two nodes is measured based on these three attributes.

B. Similarity Measures

The similarity of two nodes can be measured based on their attribute values. Here we use exact comparison of corresponding fields to compute similarity of two nodes. In fact we use Formula 6 to compute the similarity of two nodes N1 and N2.

\[
sim(N_1, N_2) = \frac{\sum \text{compare}(a_i, a'_i)}{3}
\]  

(6)

In Formula 6, \(a_i\) and \(a'_i\) are the \(i^{th}\) attributes of nodes \(N_1\) and \(N_2\) respectively. If two nodes have the same value on a given attribute then compare function returns 1 and otherwise returns 0. Of course comparing two fields can has a value in the range of 0 to 1 in general. Here for simplicity we considered it as a function with only two returned values (zero or one).

C. Implementation

We implemented the proposed method in Java with four designed classes namely PreparationGraph, Graph, Samples and SummaryGraph for this matter. We designed the first class for extracting the subgraph of each entity and put vertices and edges of that entity in separate files. Based on the figures resulted from four entities, as described in next section, we decided to select paper subgraph for summarization. Graph class has methods to construct graph, getting its vertices and edges and setting attributes of vertices. An instance of Graph class constructed for paper graph. For experimental aim, Samples class has designed to make subgraphs with different sizes of paper graph. In fact Samples class get a size as input and creates a subgraph of that size. SummaryGraph has designed to get an instance of Graph and provide methods to summarize it.

D. System Configuration

We used a system with the configuration given in Table 6 to run program and evaluate the proposed method.

<table>
<thead>
<tr>
<th>Processor</th>
<th>Intel(R)Core(TM) i7 3.5 GHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAM</td>
<td>32GB</td>
</tr>
<tr>
<td>System type</td>
<td>64 bit</td>
</tr>
<tr>
<td>OS</td>
<td>Windows 8</td>
</tr>
</tbody>
</table>

E. Results

To present an overall view of number of vertices and edges of graph, we produced 12 subgraphs that the number of vertices and edges are given in Table 7 and its chart is shown in Figure 6.

To show the order of running time of the summarization algorithm, we produced some subgraphs with sizes given in Table 8 and run summarization algorithm for each of graphs with k=10. Their running times are given in Table 8 and its chart is shown in Figure 7.

Up to now the best method to summarize a graph based on both structure and concepts is SGVR [8]. Henceforth to evaluate our proposed method, we compared our method with this method. In the aim of evaluation, we implemented our proposed method and also SGVR method and summarized graphs with

Table 7: Number of edges versus number of vertices

<table>
<thead>
<tr>
<th>#vertices</th>
<th>10</th>
<th>100</th>
<th>1000</th>
<th>2000</th>
<th>3000</th>
<th>4000</th>
<th>5000</th>
<th>6000</th>
<th>7000</th>
<th>8000</th>
<th>90000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>#edges</td>
<td>0</td>
<td>1</td>
<td>253</td>
<td>2335</td>
<td>6633</td>
<td>12526</td>
<td>19104</td>
<td>27618</td>
<td>36081</td>
<td>45888</td>
<td>57600</td>
<td>69159</td>
</tr>
</tbody>
</table>
different sizes and different values of $\alpha$ as shown in Figure 8. The summary produced with our method and with $\alpha=0.0$ is approximately equal to the summary produced with SGVR method according to Figure 8.

## Table 8: Processing times based on graph size

<table>
<thead>
<tr>
<th>Summary size (No. of vertices)</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>600</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processing time (in sec.)</td>
<td>2</td>
<td>9</td>
<td>60</td>
<td>256</td>
<td>796</td>
<td>1696</td>
</tr>
</tbody>
</table>

Density measurement has been used to evaluate the quality of produced summaries. The density of the summary graph is computed as follows:

$$\text{den}(\{V\}) = \sum_{k=1}^{k} \frac{|\{(r_p,v_q) \in r_p.v_q \cup V_1,(r_p,v_p) \in E\}|}{|E|}$$

As shown in Figure 8, the density of summary produced by our proposed method for values near to 0 for $\alpha$ is approximately equal to density of summary produced by SGVR method.
V. DISCUSSION

Our experimental results shows that the contribution of structure in the summary generated by Bei’s method is too small. As we see in columns 3 and 4 of Fig 8, the density of the summary generated based on Bei’s method is equal to density of summary of our proposed method with the structure contribution of zero. Thus Fig 8 shows that the contribution of structure in the resulting summary based on Bei’s method is close to zero. Unlike Bei’s method our proposed method is flexible and can generate a summary with any contribution degree of the structure. Summaries with different sizes are resulted because of considered different values for structure contribution in our proposed method.

The proposed method has the following features that make it superior to existing methods. In our opinion this method is the best candidate to graph summarization.

- Generate a summary with any size

In methods such as Bei’s generating a summary with any size is not possible, because in these methods an initial summary is created based on attributes and their values. For a real-world graph which every node have many attributes the initial summary is not small. But our proposed method works based on removing edges and creating a summary with a proportional size is possible.

- Producing summary based on user needs

The degree of contributions of structural and attribute similarities in the resulting summary can be determined by the user based on our proposed method. In our proposed method, it is possible to increase the contribution of the structure and attribute similarities in the resulting summary.

- Unified approach

Some methods like the one proposed by Bei at first summarize graph based on attribute similarities and then adjust the summary to support graph topology. Such methods may work for some graphs but they are
inefficient for situations where changing the resulting attributed summary to support graph topology needs a lot of vertex exchange between super-nodes. Our proposed method considers both structure and attribute concurrently and have no additional overhead.

- Learning capability

In situations where summaries are used for query answering, the contributions of structure and concepts in producing summary can be learned by the algorithm. Therefore by producing different summaries and evaluating their accuracy in answering user queries, the weighting factors of structure and concepts to constructing summary can be learned by the algorithm.

Up to now the best method for graph summarization is SGVR method [10]. As seen in columns 3 and 4 of Figure 8, our method converge to SGVR by taking the value of α near to 0. Thus our method is more flexible than SGVR method in considering any contribution degree for structure to generating summary. Because in our method the contribution of structure can be given explicitly by the user while in Bei’s method it is resulted by adjusting summary to graph structure. Thus we say that our proposed method is more general than Bei’s method.

VI. CONCLUSION AND FUTURE WORK

A new method for summarizing a graph based on both structure and concepts proposed. The proposed method implemented in Java and evaluated by real life dataset HEP-TH. We compared our method with Bei’s method [10] by implementing this method also. The experimental results showed the effectiveness of our method. The proposed method has the advantage that the contributions of structure and attribute similarities can be determined by user and for this reason generate summary based on user needs. Summary graph can be used for answering user queries. The more precisely queries the summary answers, the more the summary is better. Determining the best values for contributions of structure and concepts in producing summary is important and is one of our future plan. Learning which kind of summarization, structural, conceptual or mixture of them is necessary to answer a given query set is of our future plan. Designing a set of queries to evaluate the accuracy of summarization methods is also of our future plans.

REFERENCES


Nosratali Ashrafi Payaman received his B.Sc. degree in software engineering from Kharazmi University in 1999 and his M.Sc. degree in computer science from Sharif University of Technology in 2002. He is currently a Ph.D. student in software engineering at Iran University of Science and Technology and is also a faculty member of Kharazmi University. His current main research interests include analysis and design of algorithms, graph summarization and software vulnerability.

Mohammadreza Kangavari received his B.Sc. degree in mathematics and computer science form Sharif University of Technology in 1982, his M.Sc. degree in computer science from Salford University in 1989, and his Ph.D. degree in computer science from the University of Manchester in 1994. He is currently a lecturer in the Computer Engineering Department, Iran University of Science and Technology, Tehran, Iran. His research interests include intelligent systems, machine learning, and wireless sensor networks.
This Page intentionally left blank.
Mining Dynamic Communities based on a Novel Link-Clustering Algorithm

Hamideh Sadat Cheraghchi
Department of Computer Science and Engineering
Shahid Beheshti University
Tehran, Iran
h.s.cheraghchi@gmail.com

Ali Zakerolhosseini
Department of Computer Science and Engineering
Shahid Beheshti University
Tehran, Iran

Received: December 18, 2016- Accepted: February 27, 2017

Abstract— Discovering communities in time-varying social networks is one of the highly challenging area of research and researchers are welcome to propose new models for this domain. The issue is more problematic when overlapping structure of communities is going to be considered. In this research, we present a new online and incremental community detection algorithm called link-clustering which uses link-based clustering paradigm intertwined with a novel representative-based algorithm to handle these issues. The algorithm works in both weighted and binary networks and intrinsically allows for overlapping communities. Comparison with the state of art evolutionary algorithms and link-based clustering shows the accuracy of this method in detecting communities over times and motivates the extended research in link-based clustering paradigm for dynamic overlapping community detection purpose.

Keywords-social network; link clustering; dynamic network; evolutionary clustering; representative-based clustering

I. INTRODUCTION

So far, a large amount of research has been devoted to the task of understanding natural and hidden structures of different social, biological and information networks under the name of "graph clustering" or "community detection" approaches. In this rich area of research, communities are usually regarded as dense sub-graphs of networks. The existence of linked data is distinguishing feature of modern community detection in social network context versus traditional point-based data clustering.

Regardless of the definition, current trends in community detection present several challenges not present in traditional community detection. Dynamicity and overlapping features of the network are two key dimensions to be considered in the state-of-art algorithms due to the distinguished behavior of individuals in different social networks [1]. In the first challenge, users' joining and withdrawing behavior from communities is accounted. For the first challenge, there should be smooth changes in the discovered community, since many individuals may stay for a long time in their communities. In the case of non-overlapping community detection, this challenge is partly covered by the state-of-art evolutionary algorithms [2, 3]. However, there is always a trade-off between preserving community detection quality and smooth discovery of communities. On the other hand, there should be a mechanism to account for unforeseen dramatic shifts in the structure of communities due to the creation of new communities. The desired approach to handle this task should be computationally feasible and rapid, in order to adapt to changes. Evolutionary algorithms hardly can capture abrupt changes in the network and hence are suited for networks with small changes.
Other famous dynamic approaches such as LabelRankT [4] and QCA [5] face enormous challenge due to unpredictable behavior of networks [6]. We have recently proposed a representative-based approach called ARTISON [7] which can handle smooth community discovery while being able to handle abrupt changes in the network. Although, it is designed for non-overlapping case.

The second challenge, i.e. overlapping detection of communities in dynamic context, is a common feature of modern network. People are simultaneously members of multiple communities which create overlapping community structure. In this case, a user may hold the membership of multiple communities (e.g. family, friends, and college circles). For the case of overlapping feature, it is ideal to have solution which take intrinsically the overlapping nature of communities. Further, the challenge of handling abrupt changes still persist in overlapping context [8]. Link-clustering are the unique orthogonal solution to handle overlapping community structure intrinsically.

Our paper present a novel community detection algorithm called Link-ARTISON which target both of the challenges mentioned simultaneously. It handles abrupt changes while preserving smooth detection of communities based on the representative-based approach. Further, a novel approach in linked-based clustering is proposed which covers overlapping issue. To the best of our knowledge, simultaneous targeting of these issues is not yet covered in the literature. The contribution of the paper are as follows:

- Link-ARTISON can determine the number of communities automatically and runs using local information.
- It provides a novel dynamic link clustering algorithm by a two level linked-based similarity mechanism which can capture both low and abrupt changes.
- It outperforms the state-of-art dynamic incremental and link-based community detection in several experiments.

The paper is organized as follows: In Section II, we take a look at the background knowledge of dynamic and overlapping community detection algorithms. Section III explains our proposed approach and Section IV represents the experiments to evaluate our proposed algorithm. Finally, we present our conclusions and future directions.

II. LITERATURE REVIEW

First, Community detection in social network originally is a follow-up of a clustering approaches developed generally for data mining purpose. Survey paper [1] generally classify algorithms based on 1) Graph Partitioning methods, 2) Hierarchical clustering methods, 3) Modularity optimization methods, 4) Inference-based algorithms, 5) Spectral methods and 6) Specific approaches developed for dynamic and overlapping communities. In this Section, we concentrate more on categories of works related to our proposed algorithm and cite the recent works on each domain. Among these methods, partitioning clustering algorithms are very popular due to their simplicity and intuitiveness. In this class, a set of data points are assigned to a number of known clusters where each cluster is characterized by one prototype. Cluster are gradually constructed by iteratively mapping data points to clusters and updating representatives. One of the most popular algorithms in this category is k-mean algorithm applied in many domains. In fact, the algorithm is recognized one of the top ten influential algorithms (in 2nd place) in data mining area [9] and researchers are encouraged to further explore and find the new application and research issues of this method. The high ranking of this algorithm is due to its simplicity, intuitive optimization process and scalability properties which has paved the way for its wide application for more half a century. However, the requirement of knowing the number of clusters before computation and sensitivity to initial seeds are among the limitation of these algorithms.

On the other hand, dynamic community detection is basically categorized in two branch: 1) the first called independent community mining is a follow-up of static community discovery in each time step and comparison of the results in different steps to have a whole picture of the network. 2) The other approach called incremental community detection incorporate the information obtained in other snapshots for extracting communities of the current time step. This improves the time and computational complexity compared to independent community detection approach [10, 11]. The evolutionary clustering are an important paradigm in which a cost-function is calculated in each time step trying to minimize the changes happening to communities in the following time step. Although, this approach assumes that abrupt changes in subsequent time steps are unlikely and these changes have small impact on the community structure.

For overlapping community detection algorithms, one can recognize two broad categories of node-based and link-based algorithms that address the problem of multiple belongingness of nodes to different communities [12]. In a more detailed view [8, 13], node-based methods are subcategorized by researchers into node-seeds and local expansion algorithms like OSLOM [14], clique expansion algorithms like CPM [15] as the pioneer of overlapping dynamic algorithms, label propagation methods like dynamic overlapping SLPA[16] and other inherently dynamic and overlapping algorithms such as AFOCS[8]. In the second main category, i.e., link communities, clustering is performed on links instead of nodes [17]. This helps to provide intrinsic overlapping discovery. By finding nodes incidents to links within each community, one can recognize the structure of overlapping communities. However, the algorithms in this domain do not account for non-neighbor links which usually produce small and highly overlapping structure. Instead, we propose an approach which account for non-neighbor nodes in different iterations which works in dynamic context too.
III. Proposed Model

Let us consider a temporal network by several snapshots of the network \( \Delta = (G, G', ..., G') \). Each snapshot of the network is represented by \( G = (X', E') \). We denote the members of the network in time step \( t \) as \( X' = \{x_1, ..., x_n\} \) where each member includes its attribute including node id \((v_i)\), and weight \( w_i \) of the interaction with their neighborhood \( \Gamma(x_i) \) where \( \Gamma(x_i) = \{x_j \in X' \mid (x_i, x_j) \in E'\} \), i.e. \((w_{ij})\). Further, \( \Omega = \{C_1, ..., C_k\} \) is a family of communities (prototypes) classifying objects into \( k \) communities.

The overall procedure of the proposed algorithm are as follows: First, each node is taken iteratively and initialized by a specific link-based data structure. Then, a two-level selection process is performed to determine the assignment of the data structure to a link-based prototype structure. This two coarse and fine level process is similar to recognition and examination and their direct neighbors are stored in \( \Omega \) and the size \(|\Omega|\) of the intersection of the pair of linked input data structure \( \Phi(x_i) \) and the candidate output prototype structure \((\{'C_j|j \in \{1, ..., m\}'\}) \) and the size of the linked-data structure \( \Phi(x_i) \) under examination is calculated. The ratio is compared with the threshold value \( \delta \). The result of this equation may contribute to three cases: 1) if no similar community is found with respect to the specified threshold, a new community prototype is created to accommodate the properties of the new linked-data structure; 2) If there is an exact match for the linked-data based on eq. (1), the unique prototype is selected for inclusion of the input; 3) If there are more than one prototype selected, the fine-level selection process in the third step handles this ambiguity.

A. Initialization

First, the nodes are taken incrementally ordered by their degree. The local information of the node under examination and their direct neighbors are stored in the linked data structure of the node \( \Phi(x_i) \) as illustrated in Fig. 1. This linked data structure includes the ids of each edge, the weights and the timestamp of the interaction \( \{\{\Phi_E(x_i), \Phi_H(x_i), \Phi_T(x_i)\}\} \).

Notice that the output prototype structure store similar information to linked-data structure including ids, weights and time information. This structure includes non-adjacent nodes and links which are similar enough to each other based on the logic of the algorithm as described in the following.

\[
\Phi_E(x_i) = \{x_{i1}, x_{i2}, ..., x_{in}\}
\]

\[
\Phi_H(x_i) = \{w_{i1}, w_{i2}, ..., w_{in}\}
\]

\[
\Phi_T(x_i) = \{t, t, ..., t\}
\]

Fig. 1. The representation of linked-data structure

B. First-level link-based selection

This step involves the first process for finding the best prototype (community) for the linked-data structure. The linked-input data structure under examination \( \Phi(x_i) \) is compared with the attributes of the available prototypes \((\{'C_j|j \in \{1, ..., m\}'\}) \) according to the following equation:

\[
Sim(\Phi(x_i), C_{1,..,k}) = \frac{|\Phi_E(x_i)| \cap |C_E(j)|}{|\Phi_E(x_i)|} > \delta \quad (1)
\]

In this equation, the ratio of the size of the common members between the linked input data structure \( \Phi(x_i) \) and the candidate output prototype structure \((\{'C_j|j \in \{1, ..., m\}'\}) \) and the size of the linked-data structure \( \Phi(x_i) \) under examination is calculated. The ratio is compared with the threshold value \( \delta \). The result of this equation may contribute to three cases: 1) if no similar community is found with respect to the specified threshold, a new community prototype is created to accommodate the properties of the new linked-data structure; 2) If there is an exact match for the linked-data based on eq. (1), the unique prototype is selected for inclusion of the input; 3) If there are more than one prototype selected, the fine-level selection process in the third step handles this ambiguity.

C. Second-level link-based selection

In this second-level similarity test which aims to determine the more preferred prototype \((\{'C_j|j \in \{1, ..., m\}'\}) \) for the linked-data structure members \( \Phi(x_i) \), the similarity is assessed based on the following equation:

\[
Max(Sim(\Phi(x_i), C_{1,..,m})) = \frac{|\Phi_E(x_i)| \cap |C_E(j)|}{|C_E(j)|} \quad (2)
\]

It captures the ratio of the size of the intersection of the pair of linked input data structure and output prototype structure and the size of the given prototype under examination. Then, the candidate prototype which has the highest value of this measure is selected to hold the members of the linked data structure.

D. Update

In this step, the structure of the candidate prototype(s) is updated to reflect the changes of the selection process. There may be only an update in the weights of the prototypes or in the case of adding new members, the first attribute- i.e. the edges of the prototypes- get changed too.

Notice that for the purpose of using the past data in future join/withdraw decisions, the prototype initialization in the beginning of each time step \((t > 1)\) is committed. This is a key issue in incremental learning and compatible with the idea that members tend to preserve their membership to their old communities.
IV. RESULTS AND EXPERIMENTS

In this section, we evaluate the performance of Link-clustering on both static and dynamic real networks. The network are the famous benchmarks used for assessing the performance of the algorithms including Zachary Karate club network, Football network [19], Dolphin network [20] and Political Book [21] networks. Further, we use MIT Reality Mining [22] as a larger dataset which is typically used in the previous studies of dynamic community detection [23-25].

Since our proposed approach is a representative-based algorithm in dynamic context, the best comparison is achieved by comparing it to the state of art algorithm in representative-based category. For this reason, we choose the recent evolutionary representative-based algorithms for dynamic settings called Adaptive Evolutionary Clustering (AFFECT k-means [3]) where optimal smoothing factor is determined automatically using a statistical approach. Further, spectral algorithms are an important category of the algorithm which present high accuracy. Hence, we use the spectral version of AFFECT algorithm as another state of art evolutionary and incremental algorithm for the purpose of comparison. Further, to evaluate the strength of the algorithm in overlapping context, we use the pioneer and the most famous algorithm in link-based clustering which uses Jaccard-based similarity measure for the discovery of the communities.

For the evaluation, we use three measures to determine the accuracy of the community detection algorithms in different time steps. The widely recognized measure of Rand Index [26] is utilized for the accuracy analysis which indicates the amount of disagreement between clustering algorithm and the ground truth labels. Further, Normalized Mutual Information (NMI) - more commonly used by physicists and taken from information theory- is used to assess the quality of community detection. This measure has proven to be a robust and accurate similarity measure for a number of modalities. Finally, we take advantage of F-measure to quantify how well detected community correspond to ground truth communities. All the mentioned measures reach their best at 1 and their worse at 0 value. Further, we use overlapping NMI measure [27] specifically designed to determine the disagreement between clustering algorithm and ground truth label in overlapping context. Finally, the threshold value (δ) of the algorithm is set to 0.3 according to the experiments. Since we did not access to overlapping ground truth values of datasets, ONMI will have lower value than NMI in disjoint structure.

A. Zachary Karate Club

The well-known Karate club network shows the friendship networks of the members in this club. After a dispute between the coach and the treasurer, the club split into two clubs. The results of the comparison are presented in Fig. 4.

![Fig. 2 Performance comparison of the proposed algorithm in Zachary Karate Club network dataset with two state-of-art dynamic algorithm AFFECT k-means and AFFECT spectral in three measures of a) Rand Index, b) NMI and c) F-measures.](https://example.com/fig2)

The experiments on this network split the network perfectly into two partitions without any mismatch in different measures of Rand Index, F-measure and NMI. In all cases, the proposed linked-clustering shows much better performance. In Rand index measure, the accuracy is up to 20% better than both the state-of-art representative algorithm (AFFECT k-means) and the other dynamic spectral state-of-art algorithm (AFFECT-spectral). Further, in the case of NMI and F-measure the priority of link-clustering is more obvious which is the success of the proposed algorithm since they penalty false positive results too. In fact, the Rand index gives equal weight to false positives and false negative but F-measure penalize false negatives more strongly than false positives which is a more realistic case. Further, NMI correct the problem of comparing clusters with different number of clusters.

![Fig. 3. Illustration of Zachary Karate Club network detected by the proposed algorithm in a) disjoint structure and b) overlapping structure.](https://example.com/fig3)

The comparison of Link ARTISON algorithm with the pioneer link-based algorithm, Link Clustering [17] is presented in Table 1.

---

1 Available at [https://github.com/rabbanyk/CommunityEvaluation](https://github.com/rabbanyk/CommunityEvaluation)
ONMI measure in overlapping context also indicate almost 10% higher value than Link Clustering algorithm. Further, the number of algorithms in Link Clustering is far from the ground truth (22 cluster vs 2 clusters). Our algorithm detects the number of clusters correctly and the overlapping nodes lies in the margin between these two detected communities Fig. 3.

B. Football network

Network of American football games between Division IA colleges during the regular season Fall 2000 [19]. Edges exist if two teams played any game, and groups are conferences, scheduling groups joined by the schools for the purpose of regular season scheduling.

The results of the comparison are presented in Fig. 4. In all cases, the proposed linked-clustering shows much better performance. In Rand index measure, the accuracy is up to 20% better than both the state-of-art representative algorithm (AFFECT k-means) and the other dynamic spectral state-of-art algorithm (AFFECT-spectral). Further, in the case of NMI and F-measure the priority of link-clustering is more obvious which is the success of the proposed algorithm since they penalty false positive results too.

Since NMI in disjoint setting shows a low value of 0.19% (possibly due to mistaken ground truth as indicated by Aloise et al. [28]), the ONMI also has lower value of this measure as indicated in Table 2. However, the number of communities discovered by Link-ARTISON is much closer to ground truth communities. Both of the measures are several order of magnitude better than the results obtained by Link-Clustering. Producing a large number of highly overlapping communities is the problem common in link-based clustering. However, our algorithm does not have such problem and produce natural overlapping communities.

C. Dolphins

The other network studied to test the accuracy of the proposed algorithm is a network of frequent associations between 62 dolphins in a community in Doubtful Sound [20]. In this network, dolphins are represented as vertices, and a link is attached between two nodes if the corresponding dolphins are observed together more often than expected by chance over a period of seven years from 1994 to 2001. The groups of dolphins are mainly divided into the male ones and female ones.

The superiority of ONMI and the closeness of the number of clusters found to ground truth information is preserved in this dataset, too (Table 3).

D. Political Book

The Political book dataset [21] is a network of books about U.S. politics published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. Edges between books represent frequent co-purchasing of books by the same buyers. Groups are based on political alignment of liberal, neutral, or conservative through human evaluation.

The proposed link-ARTISON has slightly better performance in Rand Index compared to the two other evolutionary algorithm as illustrated in Fig. 6.
interactions. In this experiment, dataset recognizes a Bluetooth devices discovered during their as incoming and outgoing calls, cell tower id, and any monitoring different cell usage of participants logged approximately 500,000 hours of data is extracted by period of nine months. The large volume of consisting of students and staff interacting over a monitoring context with several distinguishing features including the intrinsic recognition of the number of communities. The initialization of each time step with previous community structure allows for smooth transition of communities between different time steps while allowing for accounting for abrupt changes in the network. The experimental results displayed the good performance of this algorithm against the state of art evolutionary algorithms and encourage the ongoing works on linked-based dynamic community detection to take more advantage of this orthogonal paradigms.

Table 4. Performance comparison of Link-ARTISON and Link-Clustering in overlapping context

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>No. of clusters</th>
<th>(Ground Truth = 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link-ARTISON</td>
<td>0.38</td>
<td>2</td>
</tr>
<tr>
<td>Link-Clustering</td>
<td>-5.2e-18</td>
<td>126</td>
</tr>
</tbody>
</table>

E. Reality mining network

The dynamic dataset of Reality Mining network analyzes the cell phone activity of 90 participants consisting of students and staff interacting over a period of nine months. The large volume of approximately 500,000 hours of data is extracted by monitoring different cell usage of participants logged as incoming and outgoing calls, cell tower id, and any Bluetooth devices discovered during their interactions. In this experiment, dataset recognizes a link between two participants by recording the IDs of nearby Bluetooth devices (student or student ID) every five minutes. Therefore, our networked data in each time step consists of two individual IDs located in nearby proximity of each other contributing to a link in network and the weight of this link is equal to the number of times these IDs are observed near each other during the time step. [22].

As the results of measures in Fig. 7 presents, the percentage of correct clustering decision reflected in Rand Index measure of link-clustering still outperforms the other three algorithm by at least 19%. For NMI measure, the trend is similar with link-clustering in the first place, AFFECT k-means in the second and AFFECT spectral in the last place of ranking. Again, link-clustering fares better than the other Evolutionary algorithms in the last two measures by almost 20%.

V. CONCLUSION

We proposed a novel linked-based incremental online community detection algorithm in social network context with several distinguishing features including the intrinsic recognition of the number of communities.

Fig. 6. Performance comparison of the proposed algorithm in Political Book dataset with two state-of-art dynamic algorithm AFFECT k-means and AFFECT spectral in three measures of a) Rand Index, b) NMI and c) F-measures.

Fig. 7. Performance comparison of the proposed algorithm in Reality mining dataset with two state-of-art dynamic algorithm AFFECT k-means and AFFECT spectral in three measures of a) Rand Index, b) NMI and c) F-measures.

REFERENCES

10. Takaffoli, M., et al. Tracking changes in dynamic information networks. in Computational Aspects of
Hamideh Sadat Cheraghchi received the BSc degree from Azad University, South Tehran branch, in 2006; M.Sc. from Azad University, Qazvin branch in 2009 and she is currently pursuing a Ph.D. degree in computer architecture in the department of Computer Engineering and Science at Shaid Beheshti University. Her current research interests focuses on data mining in social network, and soft methods under the supervision of Dr. Ali Zakerolhosseini.

Ali Zakerolhoseini received the BSc degree from the University of Coventry, UK, in 1985, MSc from the Bradford University, UK, in 1987, and PhD degree in fast transforms from the University of Kent, UK, in 1998. He is currently an assistant professor in the department of Computer Engineering and Science at Shahid Beheshti University. His research focuses on Reconfigurable device and multi classifiers. His current research interests are Data Security, Cryptography and Social Data Mining.
This Page intentionally left blank.
Pruning Concept Map to Generate Ontology

Maedeh Mosharraf
Faculty of Engineering, Electrical and Computer
Engineering Department
University of Tehran
Tehran, Iran
m.mosharraf@ut.ac.ir

Fattaneh Taghiyareh
Faculty of Engineering, Electrical and Computer
Engineering Department
University of Tehran
Tehran, Iran
ftaghiyar@ut.ac.ir

Received: December 22, 2016- Accepted: February 24, 2017

Abstract—Knowledge representation in the form of a concept map can be a good idea to categorize domain terms and their relations and help to generate ontology. Supplementing detail information to and pruning useless data from the concept map, which likes a skeleton in evolving ontology, can be semantically accomplished using the domain knowledge. In this paper, we propose a method using structural knowledge resources as well as tacit knowledge of experts to generate the ontology of eLearning domain. The concept map of eLearning is manually improved and finally verified using the group of eLearning experts. In order to enrich the ontology with merging into upcoming terms, the paper proposed an automatic method based on two external knowledge sources, Wikipedia and WordNet. The semantic similarity of concepts which is measured using the words hierarchy of WordNet combined with relations of concepts extracted from the Wikipedia graph is applied to link the new eLearning concepts to the domain ontology. The generated ontology is a dynamic knowledge source which can improve itself gradually. This integrated knowledge of eLearning domain can be used to model educational activities and to build, organize, and update specific learning resources.

Keywords—concept map; pruning; ontology generation; ontology enrichment; eLearning; graph clustering; Wikipedia; WordNet

I. INTRODUCTION

The aim of the semantic web is to enable machines to interpret and process information so that support people in doing different works on the web, especially search [1]. Several technologies that provide formal descriptions of terms, concepts, and relationships within a given knowledge base assist semantic web to its goal. Ontology is considered as one of the pillars of the semantic web technologies [2]. Although there is not a universal consensus on the precise definition of ontology, it is generally accepted that ontology is a formal specification of conceptualization [3].

Generating a worldwide ontology, which includes identifying, defining, and entering concept definitions and their relationships, is a challenging issue in the semantic web and is still far from being fully implemented. This process is so cost and time-consuming. In addition, manual process of ontology construction is limited to a special domain which requires deep understanding of that. Even in a specified domain, different opinions about concepts and their relations leads to different forms of ontologies, that none of them are sufficient certainty [4]. (Semi-)Automatic generation of ontology can overcome some of these problems.

The importance of a domain ontology is widely recognized, particularly in relation to the expected advent of the semantic web applications. The goal of a domain ontology is providing the background knowledge for any agent and function of a system and reducing the conceptual and terminological confusion among the related modules. This is achieved by the explicit representation of as more domain concepts as possible and their relationships.

E-Learning as a solution of information technology to promote educational activities provides many applications, services, resources, and systems which can benefit a domain ontology to
promote their usages. The eLearning specific ontology fosters:

- Automation of many processes in eLearning applications
- Modeling and managing different modules of eLearning systems
- Communication and cooperation among different parts of a system
- Interaction between independent systems
- Sharing and reusing educational services and resources especially open educational resources
- Profiling eLearning users as well as resources
- Development of a common language for web service interactions

This paper proposes a three-phase method for semi-automatic construction of eLearning ontology and enriching it using external knowledge bases. In the first phase of ontology generation, a hybrid method of text processing and natural language processing techniques is combined with statistical analysis to extract knowledge semantically. By applying some eLearning specific rules, the process of ontology generation focuses on this domain. This simple ontology, which is actually a concept map, is generated according to a large set of papers from a famous eLearning conference as the background knowledge. In the second phase, the generated concept map is pruned and improved to the ontology. Applying tacit knowledge of domain experts, type of each node and its relations are determined to the concept map and missing relations are added. Considering comments of all the experts, the third phase of our methodology is accomplished to enrich the generated ontology with new upcoming terms in the domain and convert the ontology to a dynamic knowledge source. Wikipedia and WordNet are used to define the meaning, appliance, and relations of new terms with the other existing concepts of the created ontology.

The rest of the paper is as follows: In section 2 a review of the related works on ontology generation methods is presented. In section 3, we propose our approach to generate the eLearning specific ontology semi-automatically. Section 4 represents the experimental results, and finally in section 5 the work is concluded.

II. RELATED WORKS

Raising interests to research about semantic web, lots of methods are proposed to generate ontology. Although a manually generated ontology is much more precise and reliable, constructing ontology (semi-) automatically is the central point of recent studies. However, it could be deficient since it relies only on pure data and not on human judgments. Typically ontology can be extracted from various data types such as textual data [5], knowledge-base [6], relational schema [7], and social networks [8]. Generating or learning ontology is the process of identifying terms, concepts, taxonomic relations, non-taxonomic relations, and optionally axioms; and applying them to construct knowledge sources [9, 5].

Reviewing (semi-) automatic ontology generation techniques, [10] groups them into four main categories: 1. Conversion or translation, which transforms the representation of an existing ontology to common knowledge representations. Conversion of XML to OWL or other ontology formats is an example. For instance, [11] develops an OWL-based language that can transform XML documents to arbitrary OWL ontologies and overcomes to shortcomings of not OWL-centric methods. 2. Mining-based methods implement some mining techniques to retrieve information and produce ontology. These techniques are usually focused on processing unstructured resources like text documents or web pages through sets of linguistic, statistical, and machine learning methods [7, 3]. Linguistic-based techniques which are mainly dependent on natural language processing tools include part-of-speech tagging, sentence parsing, syntactic structure analysis, and dependency analysis [12]. Statistic-based techniques consist of information retrieval and probabilistic patterns which provide various algorithms for analyzing associations between concepts [5]. The main idea behind these techniques is that the co-occurrence of lexical units in text often provides a reliable estimate about their semantic identity. Data mining methods can also be included in machine learning based techniques which extract rules and patterns out of massive datasets in a supervised or unsupervised manner [13]. An example of the mining-based method is [14], which benefits from the combination of C-value method, artificial neural networks, Bayesian network, and fuzzy theory to construct an ontology.3. External knowledge-bases, which build or enrich an ontology using external resources like existing ontologies, search engines [15], general knowledge resources such as WordNet [16] and Wikipedia [17]. 4. Frameworks, which provide a platform with different modules to assist ontology generation. Protégé as one of the most popular frameworks is an open source platform developed by the Stanford Medical Informatics group at the University of Stanford [18].

The other view on ontology generation methods groups them in two categories as supervised and unsupervised. Supervised methods need some training data which is labeled based on predetermined features. For example, [19] implements a tool named TextRunner which operates in three phases: in the self-supervised learning phase, a classifier is generated which labels selected words. In the single-pass extraction phase, all the relation tuples are extracted from the dataset. A probability is assigned to each tuple which is evaluated in the third phase as redundancy-based assessment. However, in unsupervised methods, hidden knowledge is extracted from unlabeled data. In the unsupervised method proposed in [20], a fuzzy version of a decision tree is used. In this research done for planning an emergency center, language predictions, categories, and describing days by activities and information about the center,
the daily working cycles for each category are identified.

III. METHODOLOGY

Domain specific ontology generation needs the strong background knowledge about that domain. However, there is not a rich knowledge source to be used in automatically generating or updating eLearning ontology. Especially in the case of new terms and concepts, related background is not rich enough to show appropriate relations. So, we suggest a three-phase ontology generation method. In the first phase, using lots of domain related documents, we extract a primary concept map consisted of frequent domain terms and relations. In the second phase, all the terms and relations of the concept map are reviewed to determine the classes, instances, and type of their relations in the ontology. Clustering the generated ontology, new terms can be gradually increased to the ontology in the third phase. Fig. 1 illustrates the detail of each phase as well as the input and output of it.

A. First phase: concept map generation

Receiving experts' opinions person to person and without intermediaries in order to find domain concepts and their relations can be so cost and time consuming. Collection of documents generated by domain experts can be an alternative for using experts' opinions and automatically generating ontology. If this ontology is supposed to be extracted from several texts, they should be numerous enough to be sure about its comprehensiveness. Our focus is on the domain of eLearning. So, we take the proceedings of ICALT (International Conference on Advanced Learning Technologies) at six years as our input corpus.

As illustrated in Fig. 1, extracting a simple ontology is accomplished in the two steps. In the pre-process step, a collection of keywords is extracted by accomplishing candidate words extraction, compound words solidification, words unification, and words standardization procedures. The set of keywords is connected in the form of a graph in the second step. In this respect, low-score words which are considered as outliers should be removed. Afterward, each pair of words which has statically potential to be linked is connected to each other by the process of edge weight calculation. Finally, applying some rules fitting the domain of eLearning, the generated graph is refined. [21] explains these steps gradually and in full detail.

Fig. 1. Three-phase ontology generation method
However, extracting ontology from the domain related corpus leads to a simple ontology which is similar to a concept map. This concept map does not provide any information about role of the concepts and type of their relations. Therefore, the transition phase for improving it to a complete ontology is needed.

B. Second phase: concept map conversion to ontology

We described that the generated graph is a simple ontology and in the other word a concept map. This concept map shows the extracted concepts of eLearning domain and their relations. Surely, types of concepts as well as kinds of edges are not specified. We can say this concept map is a skeleton for implementing the ontology. This skeleton has some weaknesses in representing the domain knowledge.

- The initial corpus that the concept map has been extracted from is a set of research papers. In these articles, with a high probability new research findings are discussed and elementary or fundamental topics are rarely explained. Therefore, there may be some fundamental concepts not covered in this collection or removed as outliers in the first phase.

- The concept map is extracted from a set of documents, so it represents the knowledge which is embodied in them. According to [22], the type of knowledge that can be codified and represents in a text document is the explicit knowledge. In this respect, we should find a method that can complete this knowledge and enrich our ontology to the other type of knowledge which is named tacit knowledge.

We should complete the ontology using the skeleton of concept map. In the other word, we should prune some useless data from the concept map and grow some details and necessary information. The steps are taken to this end are as follows.

1) Classes and instances determination

All the terms which are included in the concept map can have different roles in the ontology such as class, instance, and even property. In order to convert the concept map to the ontology, its node should be examined from this perspective and their role should be determined. Following rules make it easier.

A term is considered as a class if:
- It has a role in eLearning systems.
- It represents a resource or tool which is available for learning.
- It introduces a learning activity.
- It plays an important role in learning processes or environments.

A term is considered as an instance if:
- It is applied as an example for a class.

A term is considered as a property if:
- It introduces a feature of a class such as an element that is used for profiling or modeling.

Linguistic rules can help to find instances and properties in a text document. Phrases such as sample of, is kind of, instance of, and such as are some indications of instances in a text. However, there may be an instance applied in a text document without using these phrases. Patterns of applying a property in a text are usually in the form of class property or property of class. However, there are also many violations. Benefiting experts' knowledge, these cases can be determined. In this respect, we focus on the second phase of Nonaka and Takeuchi's organizational knowledge creation framework - called creating concepts phase ([22]).

Terms whose roles are determined, using the specified tags are introduced to the ontology. Some examples are as follows.

```xml
<Declaration>
  <Class IRI="#student"/>
</Declaration>

<Declaration>
  <ObjectProperty IRI="#age"/>
</Declaration>

<Declaration>
  <NameIndividual IRI="#MOODLE"/>
</Declaration>
```

In addition, synonym terms were unified in the concept map generation phase and replaced with a super node. Now, all of the synonyms should be added to the ontology and their relation should be determined. For example:

```xml
<EquivalentClasses>
  <Class IRI="#student"/>
  <Class IRI="#learner"/>
</EquivalentClasses>
```

2) Properties definition

As we say, many attributes of the classes are included in the concept map and are determined in the previous sub-section. These attributes are the ones that significant number of researches being accomplished on them. User characteristics are some of these attributes used in user modeling and personalization processes. Nonetheless, many features of the ontology classes are rarely considered in researches and not included in the concept map. These features may be required in various applications and future researches, so should be defined in the ontology. Using some standards improved for the domain of learning and education, such as IEEE LOM, which is improved for modeling learning objects, and SCORM, which is improved for sharing objects, can benefit in this activity. Finally, using the knowledge of experts for completing features is the additional solution.
Assigning each property to the related class is the other activity which is done through specified format and property tags of the OWL.

3) Relations labeling

According to [21], each edge of the concept map satisfies at least one of these rules.

- An edge represents the inclusion or inheritance relation of two concepts and thus forms a concept hierarchy.
- From two concepts which are linked using an edge, one of them is a tool for doing or promoting another.
- One of the concepts involved in an edge is an action in learning or eLearning process. Verbs such as assess, assign, teach, game, study, and collaborate are examples of these concepts.

However, edge types in the concept map are not specified. This is done manually and by judging domain experts. Reviewing each edge of the concept map its type, which is among sub-class, is done by, help to, do and so on, should be determined. Nevertheless, many relations in the concept map have the type is related to. This type can be a super type for all the other types. For example, a relation with the type sub-class can also be in the type is related to. So, we need to determine this kind of relations more accurately. If relations with the type is related to do not have specifically determined, it is preferred that they are pruned from the ontology. Therefore, determining the type of each relation and importing its data in the ontology, structure of the ontology can be completed.

Introducing the type of each edge to the ontology is done by calling its nodes in the format specified in OWL.

4) Missing relations determination

Although the ontology obtained from the previous sub-section is an acceptable ontology which contains all the concept map information and can be processed by machine, it is not necessarily complete. In the other word, this ontology should be completed using more details. Importing tacit knowledge of domain experts, the ontology concepts and their relations have been reviewed again and incomplete information is corrected and completed. Completing relations between concepts that are sometimes associated with adding new nodes to the ontology is an important task. The results of applying this step on the ontology of eLearning show that nearly 60 percent of added edges have the type sub-class and are completing the taxonomy of concepts.

5) Expert validation

Although a positive impact of the ontology on some applications reflects its authenticity [21], [23], we use the judgments of some experts to verify its correctness and comprehensiveness manually. In this respect, the generated ontology is sent to a group of domain experts and asked them to express their opinions about the following questions:

- Do the ontology terms cover all concepts related to eLearning?
- Does the ontology contain all relations between concepts?
- Have the type of relations been established correctly?

In this respect, we invite from seven experts of the domain to help us in this research. About all the questions, we asked the experts to express some samples which violate giving positive responses. The comments of all the experts confirm the implemented method for generating the concept map [21] and converting it to the ontology. However, considering the elimination step of node and edge outliers, some of them don't know the generated ontology as the comprehensive knowledge base. Eliminated outliers aren't justified statically; however the experts believe that they can be semantically corrected. The previous sub-section, which adds missing relations to the ontology, can compensate the missing information about ontology edges. But in the case of nodes, the ontology can be relatively weak. Therefore, we should find an appropriate solution for completing the ontology by outlier nodes and also upcoming new terms.

C. Third phase: ontology enrichment

Considering judgment of the eLearning experts, eliminating domain related outlier nodes from the concept map can blemish to the ontology. These nodes are removed due to their low frequency presence in the corpus documents. Therefore, we can say the background documents are not strong enough to statically support the ontology for adding low frequency terms. The dilemma of lacking adequate background knowledge increased when we want to merge the ontology with some terms which have been added to the domain concepts recently. However, each term has the specific semantic features which can be extracted from updated external knowledge bases.

The proposed approach uses Wikipedia and WordNet to specify application domain and semantic features of the input terms, which are removed as outliers in the concept map generation phase or recently added to the domain. Wikipedia as a knowledge base developed by collective intelligence distinguishes words with multiple meanings. Existence of a page related to each input term and following its input and output links can lead to determination of its domain. After defining the application domain, other characteristics of the
input term such as it synonyms, antonyms, parents, and grandparents in the hierarchy of words can be extracted from WordNet.

1) Ontology partitioning

Graph partitioning can facilitate the process of analyzing the structural and functional properties of the generated ontology, which is now a large and complex graph. Graph partitioning should be done semantically, meaning that the nodes placed in a group should be semantically related. Accordingly, the appropriate place for inserting new nodes to the ontology can be found locally. In this respect, graph partitioning methods can be applied to cluster the ontology. Therefore, sets of nodes should be determined so that the relation weights of any connections to nodes outside the sets are semantically higher than the relation weights of any connections to nodes inside the sets. This definition means modularity maximization [24]. After studying four algorithms, we applied a combination of label propagation [25] and Markov clustering [26] algorithms. Table 1 demonstrates the advantage and disadvantage of the investigated algorithms.

In label propagation, which is run iteratively, each node of the network is given a unique label initially. At each iteration, each node updates its label by choosing the label that most of its neighbors have. If multiple maximal labels exist among neighbors, the new label is chosen at random. The propagation iterations are performed until each node has a label that is the most frequent label among its neighbors.

Markov clustering partitions a graph via simulation of random walks. The idea is that random walks on a graph are likely to get stuck within dense sub-graphs rather than shuttle between dense sub-graphs via sparse connections. Utilizing this algorithm, the nodes in the graph are divided into non-overlapping clusters. Thus, nodes between dense regions will appear in a single cluster only, although they are attracted by different groups.

The fusion of the results obtained from label propagation and Markov clustering is performed as follows:

- If there is an overlap between the results of label propagation and Markov clustering, the common cluster would be the final cluster.
- If the result of clustering with one algorithm is a combination of other clusters from the other algorithm, then the largest cluster would be the final cluster. The smaller clusters might still exist in a hierarchy.
- If there is no overlap between two clusters obtained from two algorithms, then the cluster with maximum modularity will be the final cluster. Modularity is defined by equation 1 [24].

\[
Q = \frac{1}{2m} \sum_{vw} A_{vw} \delta(C_v, C_w)
\]

In this formula, \(m\) is the indicative of the number of edges. Let the adjacency matrix for the network to be represented by \(A\). \(A_{vw} = 0\) means there's no edge between nodes \(v\) and \(w\) and \(A_{vw} = 1\) means there is an edge between the two nodes. If we suppose the vertices are divided into clusters such that vertex \(v\) belongs to group \(c\), \((C_v, C_w)\) is defined to be 1 if two nodes \(v\) and \(w\) belong to the same group and zero otherwise. \(Q\) will be large for good divisions of the network, in the sense of having many within-cluster edges.

2) Appropriate part/s finding

It is likely that new concepts, adding to the existing ontology are related to each other. Therefore, we use an idea called Memory Cell. Memory Cells remember the situation of several last concepts which are added to the ontology. These cells cause in facing new concepts, the clusters of previous concepts are specially checked. Using Memory Cells is not possible for the first input concept. In addition, it is conceivable that input concepts are not related to each other. In order to increase the precision and avoid searching all the ontology for adding new concepts, we use a supplementary approach.

In the supplementary method, we calculate the semantic similarity of the input concepts with the delegate of each cluster in the ontology. The delegate node in each cluster can be the hub or a

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label Propagation</td>
<td>Short runtime</td>
<td>Failure to produce a unique answer</td>
</tr>
<tr>
<td></td>
<td>No need to information about the graph structure</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Propagating label of each node to its neighbors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>makes this method appropriate for clustering semantic networks</td>
<td></td>
</tr>
<tr>
<td>K-Means</td>
<td>Non-overlapping clusters</td>
<td>Need to determine the number of clusters as the algorithm input</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Considering the Euclidean Distance as similarity measure</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unsuitable for non-spherical clusters</td>
</tr>
<tr>
<td>Markov</td>
<td>High speed and scalability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Resistant to noise</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-overlapping clusters</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Considering the graph flow rather than the graph structure of the method makes it appropriate for partitioning any graph</td>
<td></td>
</tr>
<tr>
<td>Girvan-Newman</td>
<td>Focusing on edges that are most likely “between” communities</td>
<td>Long runtime</td>
</tr>
<tr>
<td></td>
<td>Inappropriate for large graphs</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 1. FEATURES OF THE CLUSTERING ALGORITHMS
node with the minimum total distance from the other cluster nodes. According to the size of the ontology and the number of its clusters, one/some of the clusters which have/have the closest semantic similarity to the input concepts is/are selected to search exactly. The semantic similarity measurement is done using equation 2.

Therefore, several clusters are suggested for each of the input concepts according to:

- Memory Cells
- Semantic similarity measurement.

3) New concepts merging

Adding the new concept to the ontology and linking it to the existing nodes are done based on their combinational tendency. Determining the threshold for combinational tendency is dependent to the domain of the ontology and can be accomplished based on experiments. For each of the selected clusters, the combinational tendency of the input concept and all the cluster nodes are calculated. Semantic features extracted from Wikipedia and WordNet are used to determine their combinational tendency.

At first, the synonyms of the input concept and all the concepts of selected clusters are extracted from WordNet. In the next step, corresponding pages of them on the site of Wikipedia are fetched. Existence of a direct link between concepts or a path with the length of two edges can connect each pair of concepts.

In our proposed approach, the dictionary of WordNet is applied when Wikipedia fails to link concepts. Failure of Wikipedia occurs in two circumstances:

- Lack of a dedicated page for each concept and its synonyms
- Lack of a direct link or a path with the length of two between each pair of concepts (or their synonyms)

Measuring the semantic distance of each pair of concepts using WordNet determines the possibility of their connection. If this distance is lower than the predefined threshold, two mentioned concepts are linked with the edge weighted by the inverse number of the semantic distance. Various methods of measuring similarity according to WordNet are introduced. [27] illustrates that Jiang and Conrath's measure is one of the best method as the only available information is the domain of concepts. The measure is one of the best method as the only available information is the domain of concepts. The measure is one of the best method as the only available information is the domain of concepts. The measure is one of the best method as the only available information is the domain of concepts. The measure is one of the best method as the only available information is the domain of concepts.

$$\text{dist}_{IC}(c_1, c_2) = 2 \log \left( \frac{\text{iso}(c_1, c_2)}{p(c_1) + p(c_2)} \right)$$  \hspace{1cm} (2)

Where \( \text{iso}(c_1, c_2) \) is the information content of the closest common concept of \( c_1 \) and \( c_2 \). In the above formula \( p(c) \) is the probability of encountering an instance of a synset \( c \) in some specific corpus.

## IV. RESULTS

As we mentioned in section 3, effectiveness of the generated concept map is evaluated through some applications [21], [23]. Table 2 represents the details of generated concept map.

### Table 2. Concept Map Characteristics

<table>
<thead>
<tr>
<th># nodes</th>
<th># edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>108</td>
<td>454</td>
</tr>
</tbody>
</table>

Applying all the activities of second phase in order to prune useless data and improve the concept map with some details followed by completing and verifying by the group of experts, the generated ontology has the specified features (Tables 3). This ontology contains 13 different relations. Since the ontology edges are two-sided, it has 26 various types of edges.

### Table 3. Ontology Characteristics

<table>
<thead>
<tr>
<th># classes</th>
<th># instances</th>
<th># properties</th>
<th># sub-class relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>171</td>
<td>51</td>
<td>86</td>
<td>152</td>
</tr>
</tbody>
</table>

The process of evaluating the third phase of ontology generation is accomplished through adding several concepts, including conceptual model, open source, Kinect, exercise, authorship, editor, agent, OER, regular, disable, MOOC, and Coursera.

Experimental results showed that the appliance of WordNet as a general purpose dictionary does not provide a good solution for eLearning domain. The main reasons are as follows:

- Various concepts in the eLearning domain are composed of multiple words and the complete form of them is not involved in the general purpose dictionary. Open source and conceptual model are some instances.
- Some domain specific words are the acronym of compound words validated only in the same domain. OER is an instance.
- Many words applied in the domain associate to special tools or methods of that domain. Kinect and coursera are placed in this group.

Considering the reasons, the usage of WordNet is beneficial in only six input terms. Table 4 indicates the number of edges added to the ontology graph for each concept.

### Table 4. The Number of Edges Added to the Ontology Using WordNet

<table>
<thead>
<tr>
<th>Concept</th>
<th>#</th>
<th>Concept</th>
<th>#</th>
<th>Concept</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exercise</td>
<td>6</td>
<td>Authorship</td>
<td>4</td>
<td>Agent</td>
<td>1</td>
</tr>
<tr>
<td>Disable</td>
<td>3</td>
<td>Editor</td>
<td>2</td>
<td>Regular</td>
<td>0</td>
</tr>
</tbody>
</table>

The type of the new edges that connect each two concepts can be in the form of is related to. However, in 39% of the new links, the edges are not reasonable in the domain of eLearning.

The encyclopedia of Wikipedia is an important source of information, so that each page is covering a title and its links are indicators of its semantic relations. One advantage of using Wikipedia is its
possibility to covering numerous titles. However, there are some problems in using this source.

- Many pages in the site of Wikipedia are linked to names or addresses of persons, organizations, or other proper names. These terms cannot be considered as classes. So, we can add them to the ontology in the role of instances.
- The title of many pages in Wikipedia is not a term or a concept. List of Latin words with English derivatives, Analysis of algorithms, List of computer scientists, Field-programmable gate array, and Talk: Computer architecture, and Scientific journal, are some instances.
- Some output links of a page are more explanations or examples mentioned for justifying the page content. Many of these links don't demonstrate a semantic relation.

Table 5 demonstrates the number of links created for each of the input concepts.

<table>
<thead>
<tr>
<th>Concept</th>
<th>#</th>
<th>Concept</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coursera</td>
<td>4</td>
<td>Disable</td>
<td>1</td>
</tr>
<tr>
<td>MOOC</td>
<td>16</td>
<td>Authorship</td>
<td>4</td>
</tr>
<tr>
<td>Exercise</td>
<td>4</td>
<td>Editor</td>
<td>2</td>
</tr>
<tr>
<td>Open source</td>
<td>4</td>
<td>Conceptual model</td>
<td>5</td>
</tr>
</tbody>
</table>

All of the created edges are from the type of is related to. Therefore, increasing the number of input concepts transforms the ontology to a concept map again. Applying domain related rules, which combine semantic and statistic features in the concept map generation phase [21], on the new added relations can delay this conversion.

In the linking process of each concept, some new terms are added to the ontology which are counted in table 6.

<table>
<thead>
<tr>
<th>Concept</th>
<th>#</th>
<th>Concept</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coursera</td>
<td>22</td>
<td>Disable</td>
<td>3</td>
</tr>
<tr>
<td>MOOC</td>
<td>5</td>
<td>Authorship</td>
<td>4</td>
</tr>
<tr>
<td>Exercise</td>
<td>5</td>
<td>Editor</td>
<td>2</td>
</tr>
<tr>
<td>Open source</td>
<td>10</td>
<td>Conceptual model</td>
<td>4</td>
</tr>
</tbody>
</table>

About half of the new terms added to the ontology are the name of persons, organizations, places, domains of education, and examples to complete the content. These terms have not been considered as the ontology classes; but can be added to as the instances. This type checking should be done manually. Therefore, in specified periods of time, the enriched ontology should pass the second-phase of ontology generation. This is because of determining the type of new added nodes and their relations.

V. CONCLUSION

In this paper, we proposed a methodology for extracting a semantic network from a corpus of documents. Pruning useless data and improving with additional details, we converted the semantic network to the ontology. Sciences are in progress, so we enriched our methodology to a mechanism that gradually promoted the ontology and added new terms and relations to it. We applied some rules specified for the field of eLearning in the creation of ontology, so this ontology is distinguished for this domain. However, the proposed method can be personalized for any other domain.

We believe that integrating the generated ontology with content and learning management systems (CMSs and LMSs) will improve their services. Therefore, future work would involve combining ontology with a CMS. Using the CMS repository, we can incrementally refine and update the ontology and consequently better annotate the archives. One application of the generated ontology is to cluster domain specific documents. Therefore, the other future directions include finding methods that combine different features and semantics from the ontology with more advanced techniques for clustering eLearning documents.

REFERENCES


I. INTRODUCTION (HEADING 1)

This template, modified in MS Word 2003 and saved as Word 97-2003 & 6.0/95 RTF for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow margins.

(Top:20mm.Bottom:20mm,Left:25mm,Right:25mm)

II. EASE OF USE

Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the file for MSW_USlttr_format.

III. PREPARE YOUR PAPER BEFORE STYLING

Space between top of the page and title of the paper has to be 85mm wide with the title centered. Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been
defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Do not mix complete spellings and abbreviations of units: \( \text{Wb/m}^2 \) or \( \text{webers per square meter} \), not \( \text{webers/m}^2 \). Spell out units when they appear in text: e.g., a few henries, not \( \ldots \) a few H.
- Use a zero before decimal points: 0.25, not .25. Use \( \text{cm}^3 \) not \( \text{cc} \). (bullet list)

C. Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation.

IV. USING THE TEMPLATE

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is \( \text{Heading 5} \). Use \( \text{figure caption} \) for your Figure captions, and \( \text{table head} \) for your table title. Run-in heads, such as \( \text{Abstract} \), will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

B. Figures and Tables

Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation \( \text{Fig. 1} \) even at the beginning of a sentence.

ACKNOWLEDGMENT (Heading 5)

The preferred spelling of the word \text{acknowledgment} in America is without an \text{e} after the \text{g}. Avoid the stilted expression \text{One of us (R. B. G.) thanks...}. Instead, try \text{R. B. G. thanks...}. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES


Note: Please submit your proposals via IJICT website at: http://jounal.itrc.ac.ir