A Novel User-Centric Method for Graph Summarization Based on Syntactical and Semantical Attributes

Nosratali Ashrafi Payaman
Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
ashrafi@khu.ac.ir
Mohammad Reza Kangavrai*
Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
kangavari@iust.ac.ir

Received: 25/04/2018 Revised: 05/09/2018 Accepted: 28/10/2018

Abstract

In this paper, we proposed an interactive knowledge-based method for graph summarization. Due to the interactive nature of this method, the user can decide to stop or continue summarization process at any step based on the summary graph. The proposed method is a general one that covers three kinds of graph summarization called structural, attribute-based, and structural/attribute-based summarization. In summarization based on both structure and vertex attributes, the contributions of syntactical and semantical attributes, as well as the importance degrees of attributes are variable and could be specified by the user. We also proposed a new criterion based on density and entropy to assess the quality of a hybrid summary. For the purpose of evaluation, we generated a synthetic graph with 1000 nodes and 2500 edges and extracted the overall features of the graph using the Gephi tool and a developed application in Java. Finally, we generated summaries of different sizes and values for the structure contribution (α parameter). We calculated the values of density and entropy for each summary to assess their qualities based on the proposed criterion. The experimental results show that the proposed criterion causes to generate a summary with better quality.

Keywords: Graph Summarization; Summary Graph; Super-node; Semantical Summarization.

1. Introduction

Graph is widely used for modeling data and their relationships. Social networks, communication networks, web graphs, biological networks, and chemical compounds are examples of data modeling by graphs. There are several applications that generate large scale and massive graphs and extensive research has been undertaken about the theory and engineering of terra-scale graphs [1]. These graphs are massive with a high growth rate. To clarify the issue, consider the statistics of Facebook users [2], which increased from one million at the end of 2004 to 1.11 billion in March 2013.

Recently proposed graph summarization algorithms [3]-[7] reduce a massive graph to a smaller one by removing details and preserving general properties. The smaller graph can be used for query answering. Of-course, these answers are not exact and may be some errors. This kind of error is acceptable since the queries are responded quickly, required by many applications.

The most real-life applications generate attributed graphs, and a summary based on both structure and vertex attributes is a critical requirement in these applications. Structural and attribute-based relationship of vertices can be considered as current and future relationships of vertices, respectively. The first one is obvious and the second one can be justified in that, based on the existing statistics of social networks, the vertices with common attribute values are probably connected. Thus generating a summary based on both structure and vertex attributes has received growing attention of the computer scientists recent years.

Although generating attribute-based summaries is not difficult and several algorithms [8] have been proposed for this purpose, generating a summary based on both structure and vertex attributes (hybrid summary) with user-determining the degrees of each is nothing less of challenge. It is obvious that the importance of structure and vertex attributes for summarization is not the same in all applications and therefore it is reasonable to consider variable weighting coefficients for them. Recently two algorithms [9],[10] have been proposed for hybrid summarization and clustering.

In graph summarization, ontology is a critical component, required to generate a high quality summary. Ontology helps to a summary in fitting with a user's needs and appropriate size. Using ontology, it is possible to explore the relationship between two attributes, determining whether they are the same, different, one subtype of each other, among other things. By involving ontology in summarization process, it is possible to drill down or roll up on the resultant summary and generate a summary of the right size. The previous summarization methods, discussed in this paper, do not incorporate ontology in the summarization process. To the best of our knowledge, no previous method is capable of generating a summary graph that can interact with both the knowledge base and the user.

In this paper, an ontology-based interactive method has been proposed for graph summarization. This method

* Corresponding Author

can be used for various types of summarization such as structural, attribute-based or hybrid.

The proposed method has a number of advantages such as generality, user-centric, knowledge-based and interactiveness nature, which makes it ideal for graph summarization. In the following, some of these advantages have been presented.

Generality: By specifying the similarity measure of two vertices, the proposed method can generate any kind of summary including structural, attribute-based or hybrid ones.

User-centric: The proposed method can generate a summary based on structure, vertex attributes or both. The importance of the structure and vertex attributes in summarization and also the significance of the attributes can be determined by the user and incorporated in the summarization.

Knowledge-based: Using a knowledge base leads to the generation of a summary of appropriate size and that is consistent with user's needs.

User-interactive: The summarization process is a supervised method in which the user can decide to stop summarization through interacting with the program.

The proposed criterion: We propose a new criterion for termination of the summarization process. We defined the proposed criterion based on the density and entropy, which shows the quality of the hybrid summary.

The rest of this paper is organized as follows. In Section 2, related works are reviewed. Section 3 is dedicated to graph summarization and related definitions. Section 4 presents the proposed method for graph summarization. The experimental results are provided in Section 5. Discussions are presented in Section 6 and finally conclusions are drawn in Section 7.

2. Related Works

In this section, we review previous works on three different types of graph summarization to discuss the main challenges of graph summarization.

2.1 Structural Summarization

Navlakha et al. [3] proposed a summarization algorithm for structural summarization where graph compression was performed by partitioning similar nodes into one group and dissimilar nodes into different groups. Edges between every pair of super-nodes are aggregated and constitute a super-edge in the summary graph. In this method, a graph is compressed with the minimum representation cost based on the MDL[1] idea. At first, they developed a GREEDY algorithm for this purpose and then proposed a RANDOMIZED version to reduce the run-time.

In [11] another method has been proposed to summarize structural graphs. In this method, the quality of a summary is guaranteed and the graph is summarized with the aim of minimizing the reconstruction errors. The authors of this paper have presented a connection between graph summarization and geometric clustering. Based on this connection, they developed a polynomial-time algorithm to compute the best possible summary of a certain size.

In [12], three distributed algorithms have been proposed to summarize large scale graphs. These algorithms are DistGreedy, DistRandom and DistLSH which differ in how they select a pair of nodes for merging (greedy, randomly, and using locality sensitive hashing theory, respectively).

Structural summarization can be used for mining frequent patterns. Chen et al. [13] proposed a method for identifying frequent patterns by producing randomized summary graphs. In fact, summary graphs rather than original graphs are mined, which are massive and time consuming. Graph summarization can also be beneficial for subgraph mining [14] and classification of aid flyers based on their property types [15].

Structural summarization can be performed using spectral graph clustering that partitions a graph based on eigenvalues and eigenvectors of the graph adjacency matrix [16]-[20]. This technique is widely used in image segmentation and social network analysis. Spectral clustering has also extensive applications in finding communities in networks [21]. It initially converts a large graph into a small one by summarization and then the resultant small summary graph is clustered by spectral clustering [22].

Graph summarization also is also used in community detection and a growing body of literature [23]-[27] has been published on this subject.

2.2 Attribute-Based Summarization

In [8], a summarization method with two novel operations has been proposed. These operations are called SNAP[2] and k-SNAP for which used for grouping nodes and summarizing attributed graphs. Attribute compatible grouping and relation compatible grouping defined by the authors of this paper. They also improved SNAP operation by proposing k-SNAP operation, where k was the summary size determined by the user.

In 2009, Zhang et al. [5] improved the k-SNAP operation by proposing the CANAL algorithm, to categorize attribute values automatically, and providing a criterion to measure the quality of a summary.

In 2008, Chen et al. [28] proposed the OLAP framework. In this framework, the cubes were created on the graph based on dimensions and measures. In the OLAP framework, a graph is summarized based on the selected attributes and input information.

2.3 Hybrid Summarization

For clustering a graph based on both structure and vertex attributes, a method was proposed in [10]. In this method, a new graph with real and virtual links is constructed for a given graph. The virtual links are added to the new graph on the account of attribute-based similarity of vertices. This new constructed graph is called the augmented graph. The similarity of two nodes is measured based on both real and virtual links in the augmented graph.

1. Minimum Description Length

2. **S**ummarization by Grouping **N**odes on **A**ttributes and **P**airwise Relations

Another method of hybrid summarization was proposed in [9]. This method first summarizes a graph based on attributes and then adjusts the summary to the graph structure by moving nodes between super-nodes.

The main challenge in the graph summarization methods is the absence of an ontology-based method to generate a hybrid summary of the attributed graph in which the α is specified by the user. The methods proposed in [9],[10] are not ontology-based and therefore unsuitable for this purpose. The proposed method resolves these challenges.

3. Graph Summarization Notion

We here present symbols and abbreviations that, have been used in Section 3.1. In Sections 3.2, 3.3, and 3.4 we will illustrate the concept of structural, attribute-based and hybrid summarization. We also present definitions of graph summarization in these sections.

3.1 Notations

In this section, the most frequently used symbols and abbreviations in this paper have been listed in Table 1.

3.2 Structural Summarization

The definition of a summary graph according to [5] is as follows:

Definition 1. (Summary Graph) Let $G = (V, E)$ be a graph and $\Phi = \{V_1, V_2, V_3, \dots, V_k\}$ be 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 Φ is $G_s = (V_s, E_s)$ where $V_s = \Phi$ and $E_s = \{(V_i, V_j) | \exists\, u \in V_i \wedge \exists\, v \in V_j \wedge (u, v) \in E\}$.

Fig. 1 shows a graph and its structural summary. As shown in Fig. 1(a), vertices a, b and c of the original graph were grouped together and made a super-node (blue one) in the summary graph (Fig. 1(b)). The summary graph has four super-nodes corresponding to four dashed ovals of the original graph along with four super-edges. For more clarity, super-nodes have the same color as their corresponding groups in the original graph.

Table 1. Symbols and abbreviations which are used in this text

Notation	*Interpretation*
G	Graph
G_S	Summary graph
c_i	Importance of i^{th} attribute
V_i	i^{th} super-node
α	Contribution of the structure in the resulting summary
Den	Density
$G_{S_{den}}$	The density of summary graph G_S
$G_{S_{ent}}$	The entropy of summary graph G_S
p_{ijn}	The percentage of vertices in super-node V_j that have value a_{in} on attribute a_i
$\text{ent}(a_i, V_j)$	The entropy of super-node V_j on attribute a_i
$sim(v_i, v_j)$	Similarity of two vertices v_i and v_j
$sim_{st}(v_i, v_j)$	Structural similarity of two vertices v_i and v_j
$sim_{si}(v_i, v_j)$	Attribute-based similarity of two vertices v_i and v_j
$sim_{si}(v_i, v_j, h)$	Similarity of two vertices v_i and $v_j$based on attribute a_h
$val(v_i, a_k)$	The value of single-valued attribute a_k on vertex v_i
$vals(v_i, a_k)$	The values of multi-valued attribute a_k on vertex v_i

3.3 Attribute-Based Summarization

To demonstrate this kind of summarization, it is necessary to define attributed graphs. The definition of an attributed graph according to [29] is as follows:

Definition 2. (Attributed Graph) An attributed graph is defined as 4-tuple $G = (V, E, \Sigma, F)$ where $V = \{v_1, v_2, \dots, v_n\}$ is a set of n nodes, $E = \{(v_i, v_j) | 1 \leq i, j \leq n \text{ and } i \neq j\}$ is a set of m edges, $\Sigma = \{a_1, a_2, \dots, a_k\}$ is a set of k attributes. Attributes of node $v_i \in V$ is denoted by $[a_1(v_i), a_2(v_i), \dots, a_k(v_i)]$ where $a_j(v_i)$ is the observation value of v_i on attribute a_j. The set $F = \{f_1, f_2, \dots, f_k\}$ denotes a set of k functions and each $f_i: V \mapsto \text{dom}(a_i)$ assigns each node $v_j \in V$ an attribute value in the domain $\text{dom}(a_i)$ of attribute a_i $(1 \leq i \leq k)$.

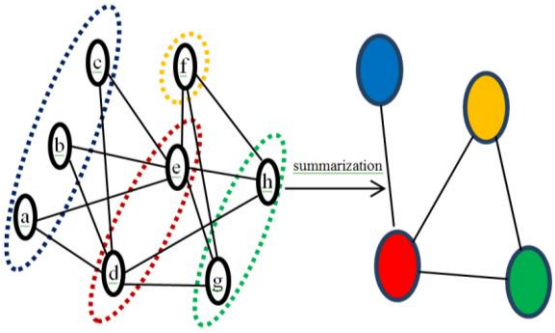

Fig. 1. (a) Original graph (left) (b) Summary graph (right)

The definition of an attribute-based summary is as follows:

Definition 3. (Attribute-based Summary) For a given graph $G = (V, E)$ let:

- Every node has an attribute set $A_v = \{a_1, a_2, \dots, a_k\}$.
- $\Phi = \{V_1, V_2, \dots, V_m\}$ is a partition on V.
- The user is interested in attributes $A_u = \{a_{i_1}, a_{i_2}, \dots, a_{i_j}\}$ where $A_u \subseteq A_v$.
- All vertices inside V_i have the same value for each attribute of A_u.

Then $G_S = (V_S, E_S)$ where $V_S = \Phi$ and $E_S = \{(V_i, V_j) | \exists u \in V_i \text{ and } \exists v \in V_j \text{ s.t. } (u, v) \in E\}$ is an attributed-based summary.

Fig. 2(a) displays an attributed graph and one of its augmented graphs is shown in Fig 2(b). In some attributed-based summarization methods of a given graph, a new graph called augmented graph is constructed. In this new constructed graph, some virtual edges are added due to the attribute-based similarity of vertices. For example, the graph shown in Fig. 2(b) is an augmented graph of the one depicted in Fig. 2(a). The weight of an edge in the augmented graph is summation of structural and attribute-based similarities of its two end vertices, but they are not necessarily equal contributions. The structural and attribute-based summaries of this graph is shown in Figs 3(a) and 3(b).

3.4 Hybrid Summarization

The definition of hybrid summarization (summarization based on both structure and vertex attributes) is as follows:

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

1. Every node has an attribute set $A_v = \{a_1, a_2, \dots, a_k\}$.
2. $\Phi = \{V_1, V_2, \dots, V_k\}$ is a partition on V.
3. The user is interested in attributes $A_u = \{a_{i_1}, a_{i_2}, \dots, a_{i_j}\}$ where $A_u \subseteq A_v$.

Then $G_S = (V_s, E_s)$ will be a hybrid summary provided that the following conditions are met:

1. G_S is a structural summary as previously mentioned.
2. All vertices inside V_i have equal value for every attribute in A_u.
3. The edge density of super-nodes is higher than a given threshold.
4. The edge density between super-nodes is lower than a given threshold.

The hybrid summary of the graph shown in Fig. 2(a) is demonstrated in Fig 3(c). The summary is generated based on both structural and attribute-based similarities. The hybrid summary as shown in Fig 3(c), is different from the other two summaries.

In the following section, the proposed method and its components are described in details.

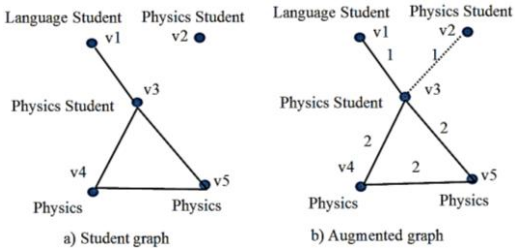

Fig. 2. Student graph and one of its augmented graphs

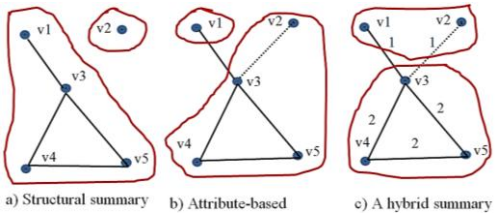

Fig. 3. Three different summaries of the student graph

4. The Proposed Method

The paradigm of the new method is shown in Fig 4. This paradigm consists of six components (four procedures, one system and criterion): 1-preprocessing 2-partitioning 3-knowledge-based reasoning 4- more summarization feasibility check 5- preparing for further summarization and 6- stopping criterion. These components have been illustrated in more details below.

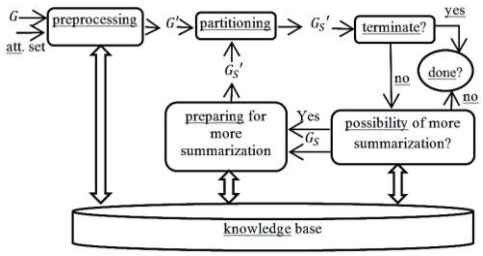

Fig. 4. The graph summarization paradigm

The new proposed method is a general one that covers three of the above-mentioned summarization. In fact, all three different kinds of summarization, only differs in terms of similarity measure, which is used merely in the partitioning component. The proposed paradigm has been summarized in Algorithm 1.

Algorithm 1: Proposed summarization method

Input: Graph G, attribute set A, importance of attributes C, summary size k, structure contribution α.
Output: summary graph G_s.

```
1  begin
2  Construct the ontology;
3  Preprocess G;
4  Partition G into super-nodes and super-edges;
5   While(Gs's size <>k or the summary
          quality is not good)
6     if(more summarization is possible)
       Change the summary for further
7        summarization;
8      Resummarize the summary;
9  end.
```

4.1 Preprocessing and Constructing the New Augmented Graph

In the preprocessing procedure, the graph and the user goal are received as the input. The user goal is expressed based on vertex attributes. The degree of an attribute's relevance to the user goal is determined by the user or by communication with the knowledge base. Unrelated attributes are removed and relevance degrees of attributes are calculated for future applications. Based on the given graph, a new graph is constructed.

The values of categorical fields and their categories or intervals are determined. Noisy values are cleansed and vertices whose errors are beyond a given threshold are removed. New attributes, which can be defined based on current attributes, are introduced and their values are calculated and stored in the knowledge base. The defined attributes represent new concepts and are useful to curtail the summary. The topics of this subsection have been summarized in Algorithm 2.

4.2 Knowledge-Based Reasoning

The knowledge base in this paradigm contains all information about nodes and their attributes, which is required for summarization. The knowledge covers concepts, individuals and their relationships and attributes values, related to nodes. In fact, the ontology of the domain is maintained in the knowledge base.

An attributed graph can contain several ontologies. For example, for business trips, there are two ontologies called geographic and financial ontologies. The geographic ontology contains information about the location of sources and destinations of trips while financial ontology describes prices, different currencies and payment methods. Most components of the proposed paradigm are engaged with the knowledge base.

Algorithm 2: Preprocessing

Input: graph G, attribute set A, importance of attributes C; summary graph G_s.
Output: the preprocessed graph

```
begin
 Remove attributes that are less related to the summarization goal;
 Remove nodes whose errors are greater than a given threshold;
 Add new required attributes;
 Categorize the domain values of numerical attributes;
 Determine the order of components for hierarchal attributes;
 Unify words by considering synonyms;
end.
```

The knowledge can be represented by semantic networks, rules and first-order logic. The first and second ones have shortcomings compared to the last one [30]. The first order logic is not suitable for this purpose due to its un-decidability. A special type of the first order logic called Descriptive logic is suitable and could be used to represent the knowledge [30]. This representation supports reasoning that obtains implicit knowledge from the explicitly stored knowledge. That is, two nodes that are dissimilar in terms of the explicit knowledge may be similar with respect to the implicit knowledge.

4.3 Partitioning

A graph is partitioned into smaller parts based on a specific similarity criterion. Vertices are grouped together based on criteria such as structural similarity, attribute-based similarity or both. In all of these three partitioning cases, the similarity of two vertices is calculated and then vertices are partitioned into smaller parts based on the similarity. In fact, similar vertices are categorized in one group and dissimilar vertices in different groups.

4.4 Stopping Criterion

Summarization process should be terminated based on a criterion. Criteria such as size and quality of summary can be used for stopping summarization process. In some cases, summarization may be stopped since further summarization is impossible. When the summary size is decided by the user, stopping criterion is obvious, obtaining a summary of that size. Otherwise, stopping criterion can be defined based on the summary quality. As the quality increases, summarization is continued. The quality of summary, depending on the type of summary, could be defined in terms of density, entropy or a combination of both as follows:

- **Structural summary:** The quality of this kind of summary is measured in terms of density. The definition of density for a summary graph with m super-nodes is as follows:

$$\text{density}(\{V_i\}_{i=1}^{m}) = \sum_{i=1}^{m} \frac{|\{(v_p,v_q)|\, v_p,v_q \in V_i \text{ and } (v_p,v_q)\in E\}|}{|E|} \quad (1)$$

- **Attribute-based summary**: In this kind of summary, the entropy measure is used to evaluate the quality of summary. The definition of entropy for a summary graph with m super-nodes and k associated vertex attributes is as follows:

$$\text{entropy}(\{V_i\}_{i=1}^{m}) = \sum_{i=1}^{k} \frac{w_i}{\sum_{p=1}^{t} w_p} \sum_{j=1}^{m} \frac{|V_j|}{|V|} \text{entropy}(a_i, V_j) \quad (2)$$

where

$$\text{entropy}(a_i, V_j) = -\sum_{n=1}^{n_i} p_{ijn} \log_2 p_{ijn}$$

and p_{ijn} is the percentage of vertices in the super-node V_j with value a_{in} on attribute a_i.

- **Hybrid summary**: The quality of a hybrid summary is measured in terms of density and entropy. In fact, Formula (3) is used for this purpose.

$$\text{Qual}(G_S) = \alpha * G_{S_{den}} / (1-\alpha) G_{S_{ent}} \quad (3)$$

Where α and $(1-\alpha)$ are contributions of structure and vertex attributes in the quality of the summary, respectively. The value of α is determined based on the importance of the structure in graph summarization.

Obviously, a good summary is the one with dense super-nodes and few interconnections. Thus, the quality of a hybrid summary is directly related to density and indirectly associated with entropy. The importance of the structure and attributes in summarization are not necessarily the same. For this reason, we multiplied density and entropy by α and $(1-\alpha)$, respectively.

4.5 More Summarization Feasibility Check

Sometimes users are interested in a summary of a specific size. Thus, the summarization process should be continued to obtain a summary of that size. However, it is not possible to summarize a graph to obtain a summary of an expected size. Hence, a criterion is necessary to check the possibility of further summarization.

For further summarize, there are a number of options such as combining intervals (ranges) of attribute values, promotion in a hierarchical attribute (e.g. student by human) or replacing a group of attributes by a newly introduced attribute (concept). The substitution of sub-types by super-types in a hierarchical attribute can be continued to reach the highest level of the hierarchy structure. Another criterion for stopping summarization process is based on the summary quality measures such as density and entropy, which are presented by Equations (1) and (2). In fact, when the summary size is not specified by the user, summarization process is stopped when the summary quality is not increased. The summary quality is defined based on a compromise between density and entropy. Algorithm 3 is used for further summarization feasibility check.

4.6 Preparing for more Summarization

The degree of summarization depends on the attribute set. By changing the attribute set or attributes, the summary size changes. Changing attributes such as promotion in a hierarchical field or combining adjacent intervals of values affects the summary. Thus,

hierarchical and categorical fields provide the best fields for changing the level of summarization. Replacing some attributes by a new attribute also increase the level of summarization. Algorithm 4 is proposed for this purpose.

5. Experimental Results

In this section, tools, datasets and computations of the proposed method and finally the results are proposed.

5.1 Gephi

We used Gephi to extract structural information and visualization of the graph. Gephi is the leading visualization and exploration software for all kinds of graphs and networks.

Algorithm 3: More summarization feasibility check

```
Input:  Summary graph G_s;
Output: A boolean value;
        begin
1       if( (at least one of the hierarchical
2           attributes is not at the highest
        level)
                          or
           (combining at least two adjacent
                  interval values is possible)
                          or
             (introducing a new attribute is
                        possible)
3         return true;
4       else return false;
5       end.
```

Gephi is an open-source and free software and its latest version (0.9.1) for Windows was used in this study. Gephi can import data to social networks also Facebook or Twitter and generate a graph and clusters.

Algorithm 3: Preparing for further summarization

```
Input:  A Summary graph;
Output: A Summary graph;
1   begin
2    let a be the less important attribute;
3    if(a is a hierarchical attribute)
       consider a higher component of this
         field;
4    else if( a is a numerical attribute)
5        decrease the number of its intervals;
6    else if(introducing a new attribute is
               possible)
7    introducing a new attribute  and add it;
8   end.
```

5.2 Dataset

We generated a graph with 1000 nodes and 2500 edges using R-Mat method and associated five attributes of age, gender, country, level of education and spoken languages to its vertices. These attributes were assigned values based on existing statistics for social networks. With the aim of obtaining graph structure information such as the number of connected components and their sizes, we developed a program for this purpose. The graph contained 185 sub-graphs of the sizes 813, 2, 2, 2 and 1. It is needless to say that the last one has the occurrence of 181. We visualized this graph using Gephi, as shown in Fig 5.

The structural features of this graph were extracted by Gephi was shown in Table 2.

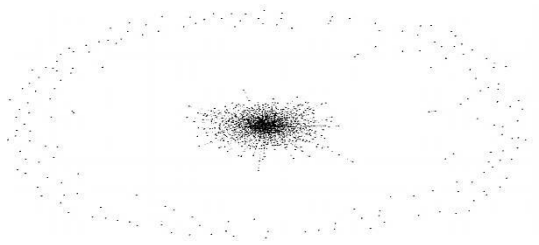

Fig. 5. Visualization of the graph by Gephi

Table 2. The extracted structural features of the graph using Gephi

average degree	diameter	Density	connected components	modularity
4.94	14	0.005	185	0.371

5.3 Computations

This section describes how to calculate the similarity of a pair of nodes, a node and a super-node or two super-nodes. That is, the necessary calculations for the hybrid summarization are presented here. The similarity of two vertices based on the structure and attributes is calculated as follows:

$$\mathrm{sim}(v_i, v_j) = \alpha \times \mathrm{sim}_{st}(v_i, v_j) + (1-\alpha) \times \mathrm{sim}_{si}(v_i, v_j) \quad (4)$$

In Equation (4), sim_{st} and sim_{si} are the structural and attribute-based similarities, respectively. These two functions are calculated as follows:

$$\mathrm{sim}_{st}(v_i, v_j) = \begin{cases} 0 & w[i][j] = 0 \\ 1 & w[i][j] = 1 \end{cases} \quad (5)$$

where w is the adjacency matrix of the given graph. If every node has k attributes, then the attributed-based similarity is calculated as follows:

$$\mathrm{sim}_{si}(v_i, v_j) = \sum_{l=1}^{k} c_l \times \mathrm{sim}_{si}(v_i, v_j, a_l) \quad (6)$$

$$\text{s.t. } 0 \le c_l \le 1 \text{ and } \sum_{l=1}^{k} c_l = 1$$

where c_l is the importance of l^{th} attribute, which it is provided by the user and $\mathrm{sim}_{si}(v_i, v_j, a_l)$ is the similarity of two vertices based on an attribute a_l which is computed as follows:

$$\mathrm{sim}_{si}(v_i, v_j, a_l) = \begin{cases} 0 & a_l \text{is single\_valued} \wedge \mathrm{val}(v_i, a_l) \neq \mathrm{val}(v_j, a_l) \\ 1 & a_l \text{is single\_valued} \wedge \mathrm{val}(v_i, a_l) = \mathrm{val}(v_j, a_l) \\ \frac{|\mathrm{vals}(v_i,a_l) \cap \mathrm{vals}(v_j,a_l)|}{|\mathrm{vals}(v_i,a_l) \cup \mathrm{vals}(v_j,a_l)|} & a_l \text{ is a multi\_valued attribute} \end{cases} \quad (7)$$

where $\mathrm{val}(v_i, a_l)$ is the value of attribute a_l on vertex v_i and $\mathrm{vals}(v_i, a_l)$ is a set of values for attribute a_l on vertex v_i. The attribute-based similarity of two vertices in terms of a multi-valued attribute is calculated based on Jaccard similarity as depicted in Equation (7).

The similarity of a node and a super-node is computed using Equation (8).

$$\mathrm{sim}(V_p, v_l) = \alpha \times \mathrm{sim}_{st}(V_p, v_l) + (1-\alpha) \times \mathrm{sim}_{si}(V_p, v_l) \quad (8)$$

The structural similarity of a super-node and a node is the number of edges between the node and the super-node divided by the super-node size. Thus, Equation (9) can be utilized for this purpose.

$$\mathrm{sim}_{st}(V_p, v_l) = \frac{|\{u|u \in V_p \text{ and } (u,v_l) \in E\}|}{|V_p|} \quad (9)$$

The attribute-based similarity of a super-node and a node indicates the summation of attribute-based similarity of the vertex and the super-node on associated attributes. Thus, Equation (10) can be used for this purpose.

$$\mathrm{sim}_{si}(V_p, v_l) = \sum_{i=1}^{k} c_i \mathrm{sim}_{si}(V_p, v_l, a_i) \quad (10)$$

Where

Table 3. Summaries with 5 super-nodes and different values of α

G_S summary	916, 47, 18, 14, 5	821, 92,54, 25, 8	558, 426, 9, 5, 24	549, 396, 50, 3, 2	415, 247, 244, 79, 15
α	1.0	0.75	0.5	0.25	0.0
density	0.4	0.107	0.230	0.172	0.123
entropy	353.947	297.09	278.950	257.926	199.914
$Qual(G_S)$	0.001	0.0003	0.0008	0.0006	0.0006

$$\mathrm{sim}_{si}(V_p, v_l, a_i) = \frac{|\{u|u \in V_p \text{ and } val(u,a_i) = val(v_l,a_i)\}|}{|V_p|} \quad (11)$$

The similarity of two super-nodes is calculated as follows:

$$\mathrm{sim}(V_p, V_q) = \frac{1}{|V_q|} \sum_{i=1}^{|V_q|} (\mathrm{sim}(V_p, v) | v \in V_q) \quad (12)$$

The quality of a hybrid summary is measured by $(\alpha * density)/((1-\alpha) * entropy)$ and it can be used as a stopping criterion in the summarization algorithm.

5.4 Implementation

The proposed method was implemented in Java for evaluation. We developed this program by designing classes such as Graph, SummaryGraph, Node, Edge, SuperNode and SuperEdge to construct and summarize a graph. The SummaryClass has a number of methods to summarize a graph and calculate the density and entropy of the generated summary.

5.5 Time Complexity

In the proposed method, the dominant time belongs to the **partitioning** component. The preprocessing and construction of knowledge base are performed once and it is also obvious that their run-times is less than the run-time of the **partitioning** component. The knowledge is stored in a tree structure. The run-time of components such as **terminate**, **possibility of further summarization** and **preparing for further summarization** are also less than the run-time of **partitioning** component. The run-time of **partitioning** is $O(n^2)$ and since this component is repeated a maximum of n times, the time complexity of the proposed method will be $O(n^3)$, where n is the number of vertices in the graph.

5.6 Results

We generated a number of summaries using the proposed method, as depicted in Tables 3, 4, and 5, to demonstrate the efficiency of the proposed method, as we did in our previous works [31] –[32]. The first row of these tables shows the size of super-nodes. For example, in Table 3, the second column of the first row indicates that the summary has five super-nodes of sizes 916, 47, 18, 14 and 5. Other rows represent α, density, entropy and the quality of each summary, respectively. As shown in Tables 3, 4 and 5, the values of density, entropy and the quality of each summary are calculated for different values of α.

To assess the quality of summaries based on the contribution of the structure in the summary, we changed the value of α from 0 to 1 with an incremental rise of 0.25 in each step. The quality of the summary based on α parameter is presented in Figures 6, 7 and 8.

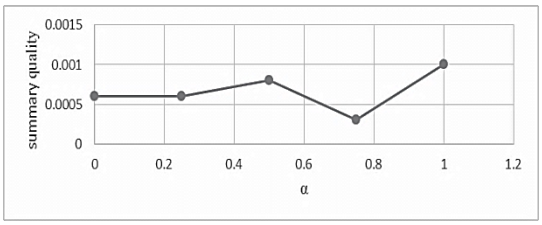

Fig. 6. The quality of summary graph in terms α parameter.

Table 4. Summaries with 10 super-nodes and different values of α

G_S summary	835, 60,40, 20,19, 13, 5, 3	818,8, 1, 3, 5, 3, 6, 3	529, 436, 15, 5, 5, 4, 4, 2	523,36, 39, 33, 5, 15, 7, 3	395,29, 181,4, 45, 3, 9, 3
α	1.0	0.75	0.5	0.25	0.0
density	0.25	0.208	0.137	0.109	0.063
entropy	315.306	297.71	269.148	241.748	182.515
$Qual(G_S)$	0.007	0.007	0.005	0.004	0.003

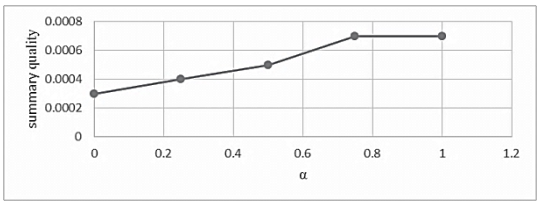

Fig. 7. The quality of summary graph in terms of α parameter.

Table 5. Summaries with 10 super-nodes and different values of α

G_S summary	830, 47, 32, 31, 31, 8, 7, 5, 5, 4	864, 68, 16, 12, 10, 8, 8, 6, 6, 2	491, 378, 33, 32, 22, 19, 14, 6, 3, 2	289, 244, 236, 134, 38, 18, 16, 15, 7, 3	472, 275, 80, 60, 52, 18, 15, 15, 10, 3
α	1.0	0.75	0.5	0.25	0.0
density	0.2	0.177	0.100	0.049	0.068
entropy	315.220	322.050	232.163	159.422	196.894
$Qual(G_S)$	0.006	0.005	0.004	0.003	0.003

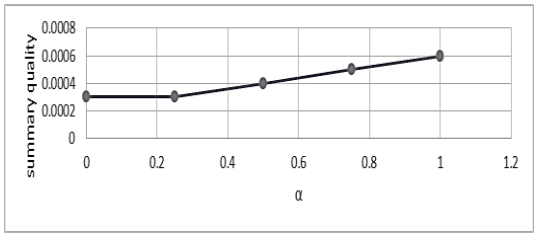

Fig. 8. The quality of summary graph in terms of α parameter.

6. Discussions

In order to assess the quality of summaries, we generated summaries of sizes 5, 8 and 10 and changed the value of α from 0 to 1 with a 0.5 increase in each step. These summaries and their features are presented in Tables 3, 4 and 5. The quality of summaries in terms of α are presented in Figures 6, 7 and 8.

Table 3 shows the summaries of size 5 with different values of α. The values of density, entropy and the newly proposed criterion, Qual(G_S), are calculated for each summary according to every value of α.

Figures 6, 7 and 8 show the quality of summaries in terms of the value of α.

As can be seen, by increasing the value of α, the value of entropy rises , but this increase is not significant and does not affect the quality of summary graph.

In these figures, by increasing the value of α the quality of summary is also improved. Hence, the quality of the summary graph is enhanced by increasing the contribution of structure in the similarity of vertices. The results suggest that the relationship of vertices in this graph is based on the connection of vertices to their similarity. In this way, we can change the value of α to generate a summary with the highest quality.

According to Figures 6, 7 and 8 and also the proposed criterion for summary quality, the best summary of the constructed graph has anα value of 1. This is in agreement with the features of the graph discovered by the Gephi tool and the program developed in Java. The best value of α, which corresponds to a summary of the high quality, can be learned by the algorithm. This is a topic to be pursued in future works.

7. Conclusions and Future Works

In this paper, we proposed a new method for summarizing a graph in an interactive and knowledge-based manner. The proposed method could summarize a graph based on users' needs. The proposed method was able to summarize a graph based on the structure, attributes or both. In this regard, ontology was used for graph summarization as it generated a summary of the right size based on user's needs. The user can determine the contributions of structure and vertex attributes in generating the summary. We proposed a criterion to measure the quality of a hybrid summary. The proposed criterion allows comparing the quality of summaries.

With the aim of evaluating the proposed method, we generated summaries of different sizes and values of α for a synthetic graph. The values of density, entropy and the proposed quality criterion were calculated for each generated summaries. To extract and visualizing the structural information of the graph, we also used the Gephi tool and an application developed in Java.

The experimental results showed that the proposed method generated a hybrid summary of higher quality. The experimental results were consistent with the graph topological structure, obtained from the above-mentioned tools.

We plan to extend the proposed method to summarize graph streams based on sliding windows to enable the monitoring of a graph stream. Using this method, hybrid summaries are compared to each other syntactically and semantically. A further research venue would be summarizing multiple graph streams.

References

[1] U. Kang, "Mining Tera-Scale Graphs: Theory, Engineering and Discoveries," 2012.

[2] "Facebook active users." [Online]. Available: https://www.yahoo.com/news/number-active-users-facebook-over-230449748.html.

[3] S. Navlakha, R. Rastogi, and N. Shrivastava, "Graph summarization with bounded error," Proc. 2008 ACM SIGMOD Int. Conf. Manag. data - SIGMOD '08, p. 419, 2008.

[4] K. LeFevre and E. Terzi, "GraSS: Graph Structure Summarization," Proc. 2010 SIAM Int. Conf. Data Min., pp. 454–465, 2010.

[5] N. Zhang, Y. Tian, and J. M. Patel, "Discovery-driven graph summarization," 2010 IEEE 26th Int. Conf. Data Eng. (ICDE 2010), pp. 880–891, 2010.

[6] M. A. Beg, M. Ahmad, A. Zaman, and I. Khan, "Scalable approximation algorithm for graph summarization," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10939 LNAI, pp. 502–514, 2018.

[7] Y. Liu, T. Safavi, N. Shah, and D. Koutra, "Reducing large graphs to small supergraphs: a unified approach," Soc. Netw. Anal. Min., vol. 8, no. 1, 2018.

[8] Y. Tian, R. A. Hankins, and J. M. Patel, "Efficient Aggregation for Graph Summarization," pp. 567–579.

[9] Y. Bei, Z. Lin, and D. Chen, "Summarizing scale-free networks based on virtual and real links," Phys. A Stat. Mech. its Appl., vol. 444, no. 2, pp. 360–372, 2016.

[10] H. Cheng, Y. Zhou, and J. X. Yu, "Clustering Large Attributed Graphs: A Balance between Structural and Attribute Similarities," ACM Trans. Knowl. Discov. Data, vol. 5, no. 2, pp. 1–33, 2011.

[11] M. Riondato, D. García-Soriano, and F. Bonchi, "Graph summarization with quality guarantees," Data Min. Knowl. Discov., vol. 31, no. 2, pp. 314–349, 2017.

[12] X. Liu, Y. Tian, Q. He, W.-C. Lee, and J. McPherson, "Distributed Graph Summarization," Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manag. - CIKM '14, pp. 799–808, 2014.

[13] C. Chen, C. X. Lin, M. Fredrikson, M. Christodorescu, X. Yan, and J. Han, "Mining graph patterns efficiently via randomized summaries," Proc. VLDB Endow., vol. 2, no. 1, pp. 742–753, 2009.

[14] S. Hosseini, H. Yin, M. Zhang, Y. Elovici, and X. Zhou, "Mining Subgraphs From Propagation Networks Through Temporal Dynamic Analysis," in 2018 19th IEEE International Conference on Mobile Data Management (MDM). IEEE, 2018.

[15] P. Pourashraf, N. Tomuro, S. B. Shouraki, "From Windows to Logos: Analyzing Outdoor Images to Aid Flyer

Classification", International Conference Image Analysis and Recognition, Springer, Cham, pp. 175-184, 2018.
[16] U. Von Luxburg, "A Tutorial on Spectral Clustering," Stat. Comput., vol. 17, no. March, pp. 395–416, 2007.
[17] I. Dhillon, Y. Guan, and B. Kulis, "A unified view of kernel k-means, spectral clustering and graph cuts", Technical Report, Computer Science Department, University of Texas at Austin, pp. 1–20, 2004.
[18] B. Auffarth, "Spectral Graph Clustering," Univ. Barcelona course Rep. Tech. Av. Aprendizaj Univ. Politec. Catalunya, pp. 1–12, 2007.
[19] S. Uw, A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," Adv. Neural Inf. Process. Syst. 14, pp. 849–856, 2002.
[20] D. Zhou and C. J. C. Burges, "Spectral clustering and transductive learning with multiple views," Proc. 24th Int. Conf. Mach. Learn. - ICML '07, pp. 1159–1166, 2007.
[21] S. Smyth and S. White, "A spectral clustering approach to finding communities in graphs," Proc. 5th SIAM Int. Conf. Data Min., pp. 76–84, 2005.
[22] J. Liu, C. Wang, M. Danilevsky, and J. Han, "Large-scale spectral clustering on graphs," IJCAI Int. Jt. Conf. Artif. Intell., pp. 1486–1492, 2013.
[23] C.-D. Wang, J.-H. Lai, and P. S. Yu, "Dynamic Community Detection in Weighted Graph Streams," Proc. 2013 SIAM Int. Conf. Data Min., pp. 151–161, 2013.
[24] G. T. Prabavathi and V. Thiagarasu, "Overlapping Community Detection Algorithms in Dynamic Networks: An Overview," Int. J. Emerg. Technol. Comput. Appl. Sci., pp. 299–303, 2013.
[25] P. Ben Sheldon, "Community Detection Algorithms: a comparative evaluation on artificial and real-world networks D . Phil student report," Other, pp. 1–27, 2010.
[26] W. Wang and W. N. Street, "A novel algorithm for community detection and influence ranking in social networks," ASONAM 2014 - Proc. 2014 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., no. Asonam, pp. 555–560, 2014.
[27] O. Benyahia et al., "Community detection in dynamic graphs with missing edges To cite this version: HAL Id: hal-01590597 Community Detection in Dynamic Graphs with Missing Edges," 2017.
[28] C. Chen, X. Yan, F. Zhu, J. Han, and P. S. Yu, "Graph OLAP - Towards Online Analytical Processing on Graphs," Data Mining, 2008. ICDM'08. Eighth IEEE Int. Conf., pp. 103–112, 2008.
[29] Y. Wu, Z. Zhong, W. Xiong, and N. Jing, "Graph summarization for attributed graphs," Proc. - 2014 Int. Conf. Inf. Sci. Electron. Electr. Eng. ISEEE 2014, vol. 1, pp. 503–507, 2014.
[30] S. Grimm, P. Hitzler, and A. Abecker, "Knowledge Representation and Ontologies Logic, Ontologies and Semantic Web Languages," Semant. Web Serv., pp. 51–105, 2007.
[31] N. Ashrafi Payaman, M. R. Kangavari, "GSSC: Graph Summarization based on both Structure and Concepts", International Journal of Information & Communication Technology Research, vol. 9, no. 1, pp. 33-44, 2017.
[32] N. Ashrafi Payaman, M. R. Kangavari. "Graph Hybrid Summarization", Journal of AI and Data Mining, vol. 6, no. 2 pp. 335-340, 2018.

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. candidate 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.

Mohammad Reza Kangavari received his B.Sc. degree in mathematics and computer science form Sharif University of Technology (1982), his M.Sc. degree in computer science from Salford University (1989), and his Ph.D. degree in computer science from the University of Manchester (1994). He is currently an Associate Professor at the Department of Computer Engineering, Iran University of Science and Technology. His research interests include intelligent systems, machine learning, and wireless sensor networks.