Open Access

Integrating Entity and Attribute for Object Similarity

Rui Xie*^,1,2^, Zhifeng Hao[2] and Bo Liu[1]

[1]*School of Automation, Guangdong University of Technology, Guangzhou, Guangdong, 510006, P.R. China;* [2]*School of Computers, Guangdong University of Technology, Guangzhou, Guangdong, 510006, P.R. China*

Abstract: In this paper, we propose a new object similarity algorithm (OSA) to address the problem of similarity calculation on the complex graph network by integrating the entity and attribute similarity of the graph node. Our proposed method can solve the similarity judgment of the objects, which are typically lack of links or attributes. Extensive experiments results have shown that our proposed object similarity method has the advantage of assessing the similarity of objects comprehensively. Our method can correct the one-sided judgment errors and improve the ability to distinguish similarity of graph objects.

Keywords: Object similarity, Entity similarity, Attributes similarity, Similarity algorithm.

1. INTRODUCTION

In recent years, many forms of networked social media, such as Microblog, Twitter, Facebook, and Wechat, have been rapidly burgeoning in terms of their membership and popularity. In the information and social networks, they may contain different kinds of attributes, and the similarity is often used to analyze the objects to find the common feature, hobbies, etc. This has led to a tremendous interest in the field of managing and mining information networks [1-4]. In addition, it can be used to cluster objects and applied in the recommendation system [5-7]. Many domestic and foreign researchers have proposed a variety of methods to measure the similarity of objects. In general, they are mainly based on semantic similarity [8] and link-based similarity [9, 10] in the homogeneous network or heterogeneous network. The problem of similarity measurement has been studied extensively in the data mining and machine learning community [11-15].

However, most of these methods are not designed for complex networks. In the complex networks, there is a certain relationship in the objects and they may have same properties. We can analyze the similarity of these objects by the similarity measure. Unfortunately, some of the objects may be missing links or properties by accident in the data collection process such as noise interference. For this reason, it is difficult for link-based measure to analyze similarity due to the lack of links or properties. For example, the Google and Baidu will not link each other directly, but they are bot h search engine companies with high similarity, since they both have the search function and provide search services. If we search a key work in both the search engines, the Google and Baidu may provide similar searching results, even more, including the items which are not provided by others. Even the two search engines have no direct link, they have high similarity with each other. Another example is that, between the Apple Company and the Xiaomi Company, there is a very tough competition between them, they do not have partnership between them, but they both have the property of High-tech information companies, such as they may produce similarity products, or they may have the same parts suppliers. In this case, the previous link-based similarity measuring methods have difficulty in calculating the similarity of objects with missing links or attributes.

This paper addresses the problem of the network with lost link or missing attributes in the object similarity measure. We propose a novel method, called the object similarity algorithm (OSA), to resolve this problem. Contrast with the previous SimRank [9] and P-Rank [10] algorithms, our method can obtain a more accurate result. Our proposed method works in three steps. In the first step, we use P-Rank method [10] to compute the entity similarity based on link. In the second step, we compute the attributes similarity based on semantics. In the third step, we integrate the entity similarity and the attributes similarity for the objects similarity. The main contribution of the paper is as follows.

(1) We propose an attributes link method and improve the attributes similarity algorithm by the attributes link, which can be well applied in the homogeneous network and heterogeneous network. We use the ratio of the attributes links and maximum number attributes of the objects; this can fully embody the characteristics of the attributes similarity.

(2) We propose the object similarity algorithm (OSA) to solve the problem of lost link or missing attributes in object similarity measure. We integrate the entity similarity and the attributes similarity for the object similarity, and effectively solve the problem of similarity measure lack of information. It can enhance the similarity of objects which is similar and reduce the similarity of objects which is dissimilar, then OSA can improve the performance of similarity computation.

(3) We compare the results of SimRank[9], P-Rank[10] with our algorithm in the experiments; the results find that

*Address correspondence to this author at the School of Computers, Guangdong University of Technology, Guangzhou,Guangdong,510006, P.R. China; Tel: 13570942092; E-mail: gdutxierui@163.com

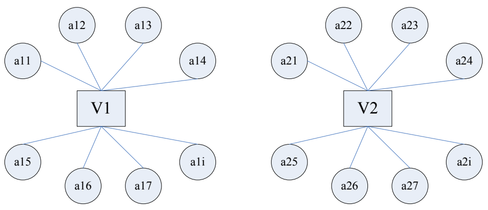

Fig. (1a). No link between entity and attributes.

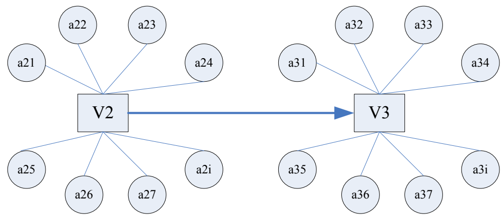

Fig. (1b). Only entity link.

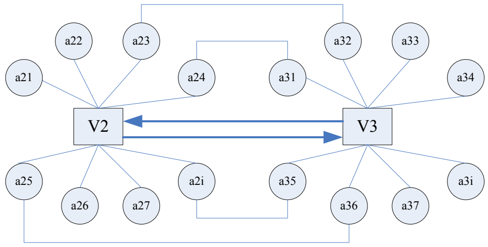

Fig. (1c). Entity and attributes links.

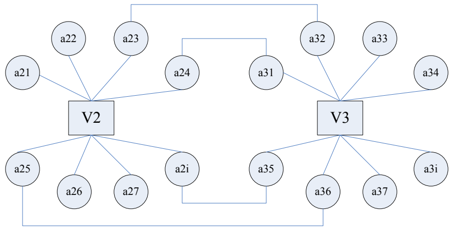

Fig. (1d). Only attributes links.

our OSA has advantage in dealing with the network lack of information, and works well compared with link-based and semantics-based methods.

The rest of the paper is organized as follows: Section 2 analyzes the entity links and the attributes links in similarity of objects; Section 3 proposes our object similarity measure method; Section 4 reports the experimental results; Section 5 concludes the paper and discusses possible directions for future work.

2. LINKS

The link information reflects the relationships between objects in graph G (V, E). In a complex network, the links represent the nature of the similarity among different types of objects. Therefore, these links are connected to objects in heterogeneous networks. Based on this, the similarity problem can be considered as a link problem in a complex network. It can be described as no links, unidirectional links and bilateral links. To make it easy to understand, we give the following definitions:

Definition 1: Entity links

In a given digraph G (V, E), the entity links means the relationships between objects, they can be classified into three categories:

(1) No Link: It means that two entities have no relationship, it can be denoted as $R(v_i,v_j)=0$(and $R(v_j,v_i)=0$).

(2) Unidirectional Link: It means that two entities have relationship but only one direction, it can be denoted as $R(v_i,v_j)=1$(or $R(v_j,v_i)=1$).

(3) Bilateral Link: It means that two entities have a bidirectional relationship , it can be denoted as $R<v_i,v_j>=1$(that is $R(v_i,v_j)=1$ and $R(v_j,v_i)=1$).

Above, R is the entity link, and V_i, V_j are the entity nodes, $R(v_i,v_j)=1$ means that the link-direction is V_i pointing to V_j.

Definition 2: Attribute links

The attributes links reflect relations between attributes of each object.

(1) In a homogeneous network, the attribute links means that two attributes of different entities have same values, it can be denoted as $L(a_{ij},a_{ji})=1$(and $L(a_{ji},a_{ij})=1$);otherwise, $L(a_{ij},a_{ji})=0$(and $L(a_{ji},a_{ij})=0$).

(2) In a heterogeneous network, the entities and the links have different types, so the attribute links mean that two different entities have same attributes, it can be denoted as $L(a_{ij},a_{ji})=1$(and $L(a_{ji},a_{ij})=1$);otherwise $L(a_{ij},a_{ji})=0$(and $L(a_{ji},a_{ij})=0$).

Above, L is attribute link, and a_{ij} is the *j*-th attribute of *i*-th entity. In the paper, the attribute is directly connected to an entity and indirect attributes belong to other entities, so the attribute links have no direction and no transitivity, it has only two cases: attribute link and attribute no link.

Thus, in a complex network, the relationships between objects include four scenarios as shown in Fig. (**1**):

3. OUR PROPOSED APPROACH: SIMILARITY CALCULATION WITH LOSING LINKS OR MISSING ATTRIBUTES

The complex network consists of many different types of objects and relationships. Because every object has different properties, we can describe an object as a node entity and its properties as node attributes. Then the complex network will be consisted of nodes, attributes and their relationships. The relationships between the entities reflect the structure of complex network and it is the global character for the complex network. The attributes represent the specific characteristics of the nodes, so it is the local character for the complex network. If the two objects are similar, then there will be a certain similarity reflected in the nodes, attributes, or relationships. It has many measures to judge similarity by semantics and links. The similarity can be analyzed by entities and attributes, or measured by links between objects. To analyze the similarity in this paper, we introduce two similar concepts of object entity similarity and attributes similarity.

Definition 3: Entity Similarity

According to P-Rank, the two entities are similar if they are related to similar entities. The meaning of P-Rank is elaborated as [10]:

1. Two entities are similar if they are referenced by similar entities.

2. Two entities are similar if they reference similar entities.

Then the entity similarity, which is called SimE (V_i, V_j), can be computed by the links of the entities. It reflects the relationship of various objects in the complex network, and the entity similarity is determined by the structure of the complex network. Hence it describes the global similarity of the objects in the complex network. The links in objects will have different semantics in the homogeneous and heterogeneous networks. Since the objects can dynamically join in and exit out a complex network, thus the links will dynamically change according to the nodes, and the entity similarity will also be changed promptly. That means the global similarity is dynamic. To analyze the dynamic of the entity similarity in a complex network, we can calculate similarity in a regular or irregular time. In this paper, we consider the complex network in a steady state at time t, and then we can calculate the global similarity of the complex network by link-based measure such as P-Rank.

Definition 4: Attribute Similarity

The attribute similarity describes the similarity of two objects by the attributes. The attributes are the characters, states or parameters of the object. If two objects are similar, the similarity will be exhibited by the attributes. Particularly, the ratio of the same attributes to the maximum attributes of one object is calculated, which reflects the degree of similarity between objects by attributes. In other words, we can use the ratio to measure the attribute similarity; the higher the ratio, the more similar the objects. And the attribute similarity can be expressed as shown in equation (1):

$$\mathrm{SimA}(a_i,a_j)=\frac{|\,a_j\ \mathrm{I}\ \ a_j\,|}{\max(|\,a_i\,|,|\,a_j\,|)} \tag{1}$$

where, a_i and a_j denote the attributes of *i*-th object and the *j*-th object. And | $a_i \cap a_j$ | denotes the number of the same attributes of two objects. The denominator is the maximum of attributes in two objects, thus 0≤SimA≤1. If two objects do not have the same attributes then SimA=0; if two objects have all the same attributes then SimA=1.

For the attribute similarity, it reflects the local similarity of objects in a complex network, and this similarity can also judge whether two objects belong to the same class or not. On the one hand, in the homogeneous network, all nodes from one class which have same attributes, the attribute similarity is the ratio of the number of the same attributes values to the total number of attributes in one node; On the other hand, in the heterogeneous network, all nodes from different classes which have different attributes, thus the attribute similarity is the ratio of the same attributes to the maximum attributes of one object. Compared to the dynamic characteristics of entity similarity, the attribute similarity is relatively static. In a graph G (V, E), V is the entity and E is the edge between the entities, thus R is the relation between entities, and A is the attributes of entities. We can conduct the attribute similarity algorithm that is called ASA as follows:

Algorithm 1: Attribute Similarity Algorithm--- (ASA 1)

```
   Input: A direct graph G (V, E) with entities, links, attributes
   Output: A similarity matrix of attribute---SimA (*,*)
   Initialize
   For each $E_i$ □G do
   For each $E_j$□G do
       If L ($a_{ij}$ ,$a_{ji}$) =1 then N=N+1
M=max (|$A_i$|, |$A_j$|)
R=N/M
SimA ($a_i$, $a_j$) =R
  Until all the entities have been computed
Return SimA (*,*)
```

It is found that we can use the link-based measure to compute the entity similarity, because it has considered the relative independent relationships of the objects and got the structure information of the complex network. In addition, it ignores the influence of the semantic to understand the complex network. Otherwise, it is easy to understand the semantic of the objects, because the attribute similarity is based on contents of the text vector space. Both methods have some advantages in similarity analysis based on unilateral aspects. This paper focuses on the objects similarity with the entity similarity and attributes similarity.

3.1. Similarity

The complex network is composed of different types of objects which have different types of relationships and attributes. Links may exist between objects, or links may not exist, and attribute values may or may not be the same in the homogeneous network. However, in the heterogeneous network, objects usually have different attributes. Therefore, it is difficult to get all the relations and attributes information in a real complex network. Thus, it is difficult to measure the similarity accurately. In order to comprehensively measure object similarity, we introduce the entity relationship model which is widely used in the database area. The object similarity is integrated based on entity and attributes similarity in our method.

We treat the objects as the entities, the corresponding properties of the objects are considered as the attributes belonging to entities, the relationships of objects are treated as the relations of entities. Then we can establish the entity relationship of object similarity model G (V, E, R, A) to measure the similarity.

3.2. Similarity Measure

In the G(V,E,R,A) model, we can separately compute the similarity of entities that reflects the global similarity, and calculate the similarity of attributes that reflects the local similarity. Considering the cases of missing links or attributes lost, it is important to balance the global similarity and local similarity to get the objects similarity. We let λ denote the weight factor of the global similarity and (1-λ) denote the weight factor of the local similarity. We can change the coefficient of both similarities to adjust their weights on the objects similarity. Then we can get the following expression for the objects similarity, as shown as equation (2):

$$\mathrm{Sim}(a,b)=\lambda\times\mathrm{simE}(a,b)+(1-\lambda)\times\mathrm{simA}(a,b)\ (0\leq\lambda\leq1)\quad(2)$$

Where Sim (*a, b*) represents the similarity between object *a* and object *b*. Sim (*a, b*) = 0 means two objects are completely different, and Sim (*a, b*) = 1 means two objects are the same object. As discussed before, λ is the entity similarity (global similarity) coefficient which reflects the entity's contribution to the object similarity, and (1-λ) is the attributes similarity (local similarity) coefficient that reflects the attribute's contribution to the object similarity. How to select the value of λ is important to measure the object similarity. In general, the two weight factors are equally important, it can be taken as λ=0.5; in some special cases, the entity similarity is more important than the attribute similarity, then λ>=0.5, otherwise λ<0.5; we can also use the sufficient information of links or attributes to select the value of λ. If links information is not sufficient, then λ<0.5, otherwise λ>=0.5. We used λ=0.5 in the experiment.

3.3. Object Similarity Algorithm

The object similarity integrates the entity similarity and attributes similarity, which includes link-based information and semantic-based information. Then the object similarity takes the advantage of local similarity and global similarity in the complex network. We can use P-Rank algorithm to compute the entity similarity and use ASA to compute the attributes similarity separately, then according to Equation (2), the objects similarity algorithm can be expressed as follows:

Algorithm 2: Object Similarity Algorithms---OSA 2

```
Input: complex network G, weight (V, E, R, A), factor λ=0.5
Output: The similarity matrix of objects--- Sim (a, b)
For each a, b□ G do / * initialize * /
If a == b then Sim (a, b) = 1
Else Sim (a, b) = 0
For each a, b □G do
/ * P-Rank algorithm compute entity similarity * /
SimE (a, b) = P-Rank (a, b)
/* ASA compute attributes similarity*/
SimA (a, b) = ASA (a, b)
Sim (a, b) *= λ × SimE (a, b) + (1-λ) × SimA (a, b)
Sim (a, b) = Sim (a, b) *
Return Sim (*, *)
```

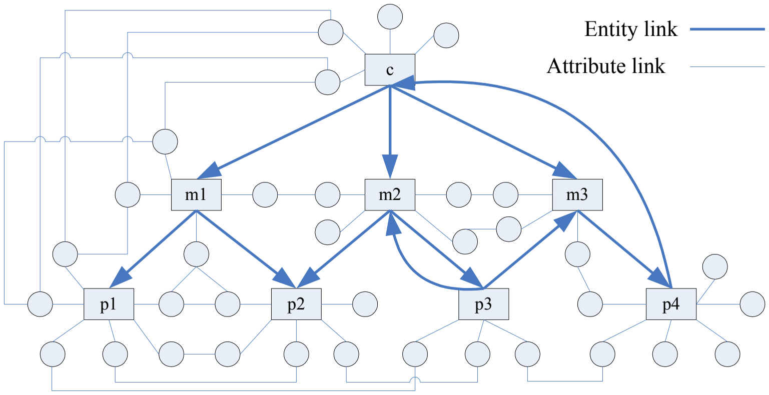

Fig. (2). The experiment graph.

Table 1. The computation result.

Object Pair	SimRank	P-Rank	OSA
{c,c}	1	1	1
{m1,m2}	0.4	0.42	0.335
{m1,m3}	0.4	0.38	0.190
{m2,m3}	0.4	0.295	0.3975
{p1,p2}	0.56	0.283	0.3915
{p1,p3}	0.32	0.176	0.1713
{p3,p4}	0.32	0.124	0.1453

The algorithm needs n^2 spaces to store the similarities of objects in the object similarity method, it is also required for the entity similarity algorithm and the attributes similarity algorithm, and finally it needs 3* n^2 space together to store the three similarity matrices. Then the space complexity is O (n^2); the worst time complexity of P-Rank algorithm is O(n^4) in the entity similarity computation process, and the worst time complexity is O (n^4) in the attributes similarity. So the worst time complexity of the object similarity algorithm is O(n^4) as well, therefore it is not suitable for large-scale network computing.

4. EXPERIMENTS AND RESULT ANALYSIS

To compare with the result of SimRank and P-Rank methods, we use the data of SimRank [9] and P-Rank [10] in our experiments, and we add the attributes links into objects. Then we use SimRank, P-Rank and our proposal OSA to compute the similarity of Fig. (**2**) which is used in SimRank and P-Rank.

We can get the result of similarity of Fig. (**2**) as shown in Table (**1**):

It has obvious influence on the similarity after the introduction of the attributes similarity to the object similarity. From the experimental results, we can find that : (1) For the same object's self-similarity, the three algorithms give the same similarity, such as the data {c,c}; (2) The object similarity has the same ability as P-Rank so that it can distinguish the similarity as SimRank, such as data {m1, m2}, {m1, m3}, {m2, m3}, and data {p1, p3}, {p3, p4}. (3) The object similarity has enhanced the ability to distinguish the similar objects. This means that the similar objects will reveal their similarities not only by links but also by attributes, and the two factors will balance and restrict each other. On the one hand, if the two objects have much similarity by links but less similarity by attributes, then we can conclude that the two objects are not really similar in some aspects, such as data {m1, m2}, {m1, m3}, and data {p1, p3}. On the other hand, if the two objects have no links that means they have no entity similarity, but they have attributes similarity, then we can compute their object similarity and infer that they lost links due to some reasons, such as data {m2, m3}. So the object similarity enhances the similarity of objects which are similar originally and reduces the similarity of objects which are dissimilar, then it can correct the inaccuracy of the similarity obtained from only one aspect.

CONCLUSION

In this paper, the object similarity has taken into account the entity similarity and the attributes similarity. It has avoided the partiality of the similarity measure as discussed before, and provided a good solution for some objects which is lack of links or attributes information caused by accident or intention. Our proposed method can be applied in both homogeneous and heterogeneous networks. In the future, we intend to improve the algorithm to reduce the complexity and apply it in the large-scale network.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is financially supported by Natural Science Foundation of China (61100148,61202269), Foundation for Distinguished Young Talents in Higher Education of Guangdong, China (LYM11060), Natural Science Foundation of China (61472090, 61203280,61202270), Guangdong Natural Science Funds for Distinguished Young Scholar (S2013050014133), Science and Technology Plan Project of Guangzhou City(12C42111607, 2014J4100201，201200000031, 2013y2-00034), Specialized Research Fund for the Doctoral Program of Higher Education (20134420110010), and Discipline Construction and Quality Engineering of Higher Education in Guangdong Province (PT2011JSJ).

REFERENCES

[1] Yingfang Li, Keyun Qin and Xingxing He, “Some new approaches to constructing similarity measures”, *Fuzzy Sets and Systems (FSS)*, vol. 234, pp. 46-60, 2014.

[2] Segla Kpodjedo, Philippe Galinier and Giuliano Antoniol,“Using local similarity measures to efficiently address approximate graph matching”, *Discrete Applied Mathematics (DAM)*, vol. 164, pp. 161-177, 2014.

[3] Chengjun Liu, “Discriminant analysis and similarity measure”, *Pattern Recognition (PR)*, vol. 47, no. 1, pp. 359-367, 2014.

[4] Francisco Chiclana, J. M. Tapia García and M. J. del Moral, Enrique Herrera-Viedma,“A statistical comparative study of different similarity measures of consensus in group decision making”, *Inf. Sci. (ISCI)*, vol. 221, pp. 110-123, 2013.

[5] Yasmina Bashon, Daniel Neagu, and Mick J. Ridley,“ A framework for comparing heterogeneous objects on the similarity measurements for fuzzy, numerical and categorical attributes”, *Soft Comput. (SOCO),* vol. 17, no. 9, pp. 1595-1615, 2013.

[5] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry, “Using collaborative filtering to weave an information tapestry”, *Communications of the ACM*, vol. 35, no. 12, pp. 61-70, 1992.

[6] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl, “GroupLens: Applying collaborative filtering to Usenet news”, *Communications of the ACM*, vol. 40, no. 3, pp. 77-87, 1997.

[7] Upendra Shardanand and Pattie Maes,”Social Information Filtering: Algorithms for Automating "Word of Mouth"”, *In Proceedings of the Conference on Human Factors in Computing Systems*, pp. 210-217, 1995.

[8] B. Saleena and S. K. Srivatsa, “ConSim: an enhanced semantic similarity measure to find the relationship between concepts in cross ontology”, *CUBE*, pp. 303-307, 2012.

[9] G. Jeh and J. Widom. “SimRank: a measure of structural-context similarity”, *In KDD'02*, pp. 538–543, 2002.

[10] Peixiang Zhao, Jiawei Han and Yizhou Sun: P-Rank: a comprehensive structural similarity measure over information networks. *Proc. 2009 ACM Conf. on Information and Knowledge Management (CIKM'09), Hong Kong, China,* pp. 553-562, 2009.

[11] Leman Akoglu, Hanghang Tong, Brendan Meeder, and Christos Faloutsos, “PICS: Parameter-free Identification of Cohesive Subgroups in Large Attributed Graphs”. *SDM*, pp. 439-450, 2012.

[12] Yizhou Sun, and Jiawei Han, “Mining Heterogeneous Information Networks: Principles and Methodologies”, *Synthesis Lectures on Data Mining and Knowledge Discovery*, Morgan & Claypool Publishers 2012.

[13] Yizhou Sun, and Jiawei Han, “Mining heterogeneous information networks: a structural analysis approach”, *SIGKDD Explorations* vol. 14, no. 2, pp. 20-28, 2012.

[14] Xiao Yu, Yizhou Sun, Brandon Norick, Tiancheng Mao, and Jiawei Han, "User guided entity similarity search using meta-path selection in heterogeneous information networks", *CIKM*, pp. 2025-2029, 2012.

[15] B. Saleena and S. K. Srivatsa, "ConSim: an enhanced semantic similarity measure to find the relationship between concepts in cross ontology", *CUBE*, pp. 303-307, 2012.

Received: May 26, 2015 Revised: July 14, 2015 Accepted: August 10, 2015