PathSimExt: Revisiting PathSim in Heterogeneous Information Networks

Leong Hou U.*, Kun Yao, and Hoi Fong Mak

Department of Computer and Information Science, University of Macau, Macau
{ryanlhu,leomak}@umac.mo, yaokun527@qq.com

Abstract. Similarity queries in graph databases have been studied over the past few decades. Typically, the similarity queries are used in homogeneous networks, where random walk based approaches (e.g., Personalized PageRank and SimRank) are the representative methods. However, these approaches do not well suit for heterogeneous networks that consist of multi-typed and interconnected objects, such as bibliographic information, social media networks, crowdsourcing data, etc. Intuitively, two objects are similar in heterogeneous networks if they have strong connections among the heterogeneous relationships. PathSim is the first work to address this problem which captures the similarity of two objects based on their connectivity along a semantic path. However, PathSim only considers the information in the semantic path but simply omit other supportive information (e.g., number of citations in bibliographic data) . Thus we revisit the definition of PathSim by introducing external support to enrich the result of PathSim.

1 Introduction

A heterogeneous network is a logical network that usually consists of a large amount of multi-typed and interconnected components. The inter-connections in the heterogeneous networks often indicate different kind of relations, such as bibliographic networks, epidemic network, and social media network [1]. As an example in bibliographic networks, users may be interested in querying similar authors to the author of a paper that they just read. There are variant ways to measure the similarity of two authors in a bibliographic network. For instance, two authors is similar if their papers co-appear in the same venue frequently or they are the co-authors in many publications.

Similarity queries in homogenous networks have been extensively studied over the past few decades, where random walk based approaches (e.g., P-PageRank [2] and SimRank [3]) are the representative methods in this category. However, the random walk solutions cannot be used in heterogenous networks since the walks over different relationships have different meanings behind. To address the similarity queries in heterogenous networks, Yizhou Sun et al. [4] proposed PathSim that computes the similarity of two objects based on the affiliation of a semantic path. For instance, the affiliations in bibliographic networks may include *authors-venues* (**AV**), *authors-papers* (AP), and

[*] This work was partially supported by MYRG109(Y1-L3)-FST12-ULH from UMAC Research Committee.

F. Li et al. (Eds.): WAIM 2014, LNCS 8485, pp. 38–42, 2014.

papers-terms (**PT**) relationships. A possible semantic path is *authors-venues-authors* (**AVA**) which indicates the similarity of the authors based on their co-appearance in the same venue. Given a semantic path (e.g., **AVA**) and a query object (e.g., an author), PathSim returns the most similar objects based on the affiliations in the semantic path.

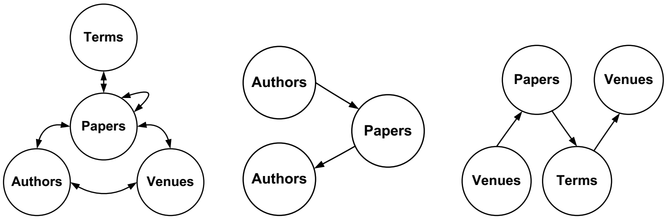
Table 1. Similarity search result under meta-path 'VAV' of a query 'PKDD' on DBLP dataset

Rank	P-PageRank	SimRank	PathSim	PathSimExt
q	PKDD	PKDD	PKDD	PKDD
1	KDD	Local Pattern Detection	ICDM	PAKDD

In this work we enrich PathSim by introducing *external support* into the similarity computation. We observe that some factor can be used as an external support for improving the similarity search. More specifically, the similarity of venues is not only reflected by the semantic path *venues-authors-venues* (**VAV**) but also the reputation similarity of the venues (e.g., average citations per paper). Table 1 compares this idea with other similarity methods using a DBLP example that lists the most similar venues to 'PKDD' based on a semantic path *venues-authors-venues* (**VAV**). 'PKDD' is a European conference on machine learning and knowledge discovery in databases. As shown in the table, P-PageRank [2] and PathSim [4] return 'KDD' and 'ICDM', respectively; however, these results may be not the best answer for the query since the reputation of 'KDD' and 'ICDM' is quite different from 'PKDD'. SimRank returns a less related conference 'Local Pattern Detection' as the top-1 result which manifests the problem of random walk based solutions in heterogenous networks. Thus, we revise the definition of PathSim, named as PathSimExt, that additionally introduces *external support* into the similarity measures.

2 PathsimExt

A heterogenous information network is a special type of information networks that either contains multiple types of objects or multiple types of relationships. More specifically, a heterogenous information network is a directed graph G, where each vertex belongs to one particular object type T and each edge belongs to one particular semantic relationship R.



(a) Semantic relationships

(b) AVA

(c) VPTV

Fig. 1. Bibliographic semantic relationships and 2 semantic paths

Semantic paths, $\varDelta$. Based on the semantic relationships, we can construct a semantic paths $\varDelta$ that typically represents some semantic meaning behind. For instance, a semantic path *authors-papers-authors* ($\mathbf{APA}$) (shown in Figure 1(b)), that uses two semantic relationships $\mathbf{AP}$ and $\mathbf{PA}$, indicates the co-authorship. Note that the semantic path is not necessary symmetric, where a asymmetric semantic path $\mathbf{VPTV}$ can be found in Figure 1(c). To facilitate reasonable search queries, we assume that the begin and the end object type in a semantic path are always identical in this work.

PathSim. As studied in [4], PathSim is shown to capture better semantics of peer similarity in heterogenous network than other methods proposed for homogenous information networks, such as P-PageRank and SimRank. Formally, PathSim is defined as follows.

$$S_{\varDelta}(x,y) = \frac{2 \times w(x,y)}{w(x,x) + w(y,y)} \tag{1}$$

where $w(x,y)$ represents the number of path instances between x and y under the semantic path $\varDelta$. The denominators, $w(x,x)$ and $w(y,y)$, can be viewed as the normalization factor in the definition.

Table 2. Search result under $\mathbf{APA}$ on a query 'Mike'

(a) Co-authorship and author information

	#collaborations $w(x,y)$	#publications $w(x,x)$	#cititations $ext(\mathbf{A},x)$
Mike	-	100	300
Jim	20	20	50
Tom	19	100	300

(b) Similarity value between Mike and others

	Jim	Tom
PathSim	0.33	0.19
PathSimExt	0.028	0.095
PathSimExt (norm.)	0.057	0.19

Table 2 shows an example of PathSim that finds the most similar authors of 'Mike' under the semantic path $\mathbf{APA}$. According to the definition of PathSim (Equation 1), 'Mike' is similar to 'Jim' than 'Tom' since the number of path instances (i.e., the number of collaborations) between 'Mike' and 'Jim' is higher than 'Mike' and 'Tom'. However, if we take the fame of the authors (i.e., number of citations) into consideration, 'Tom' should be a more similar result to 'Mike' since both of them have similar number of citations and publish certain amount of papers together. This example shows that external support (i.e., number of citations) can enrich the result of PathSim but is not considered in [4].

External Supports. An *external support* can be any factor that reflects the importance of the objects from global views. Typically, for each type of object $\alpha \in T$, we can facilitate some support (e.g., well accepted knowledge and common sense) to rank the objects $o_i \in O$ in a reasonable way, which is defined as $ext(\alpha, o_i)$ in this work. For instance, in the bibliographic network, we can use the citation numbers as the external support for every author a_i in the author type $\mathbf{A}$, which is denoted as $ext(\mathbf{A}, a_i)$. Given the external support, we revise PathSim to PathSimExt as follows.

$$S^{E}_{\varDelta}(x,y) = \frac{|w(x,y)| \times sim_T(x,y)}{ext(T,x) + ext(T,y)} \tag{2}$$

where $sim_T(x, y) = \frac{min(ext(T,x),ext(T,y))}{max(ext(T,x),ext(T,y))}$ can be viewed as the similarity of two objects in the object type T. The denominators in Equation 2, $ext(T, x)$ and $ext(T, y)$, can be viewed as the normalization factor.

3 Experiments

In this paper, we used a DBLP citation network dataset downloaded from arnetminer.org that contains 7.7K venues, 1M authors, 1.6M papers, 34K terms, and 2.3M citations.

Matrix	Data size	Density
AV	2,988,422	0.0374%
AP	4,227,433	0.00025%
VP	1,632,442	0.013%
VT	3,741,075	1.42%
VAV	7,422,651	12.5%

Rank	AVA	APA	VAVTV	VAV
q	Jiawei Han	Jiawei Han	ACM Trans. Graph.	ACM Trans. Graph.
1	Philip S. Yu	Philip S. Yu	SIGGRAPH	SIGGRAPH
2	Christos Faloutsos	Jian Pei	IEEE Visualization	Trans. Vis. Comput. Graph.
3	Divesh Srivastava	Charu C. Aggarwal	Trans.Vis.Comput.Graph.	Journal of Computer Vision
4	H. V. Jagadish	ChengXiang Zhai	ICCV	Symp. on Comput. Geometry
5	Surajit Chaudhuri	Laks V. S. Lakshmanan	CVPR	Computer Aided Geometric Design

We show the result of different semantic paths in Table 3, including author-to-author relationships (**AVA** and **APA**) and venue-to-venue relationships (**VAVTV** and **VAV**). The result of **AVA** shows that 'Philip S. Yu' and 'Christos Faloutsos' are the two most similar authors to 'Jiawei Han'. The result is meaningful as both 'Philip S. Yu' and 'Christos Faloutsos' publish lot of papers in the common venues of 'Jiawei Han' and their citation records (i.e., external support) are strong. In another set of experiments, we evaluate the path **VAVTV** and **VAV** for a journal 'ACM Transaction on Graphics'. The results using the semantic path **VAVTV** are more relevant to the area of computer graphics and computer vision research since the returned venues should have similar topics according to **VTV** (i.e., the terms). In addition, 'ICCV' and 'CVPR' are ranked in the top-5 result since these venues have similar reputation (i.e., by external support) to 'ACM Transaction on Graphics'. For clarity, the results using the semantic path **VAV** are more diverse since the path only take author-venue relationship into consideration.

4 Conclusions

In this work, we revise the definition of PathSim that enriches the result of PathSim by introducing the external support into the similarity measure. Our result demonstrates that the external support is effective in bibliography data. In the future, we attempt to further improve the query performance in terms of both efficiency and effectiveness.

References

1. Han, J.: Mining heterogeneous information networks by exploring the power of links. In: Gavaldà, R., Lugosi, G., Zeugmann, T., Zilles, S. (eds.) ALT 2009. LNCS, vol. 5809, p. 3. Springer, Heidelberg (2009)
2. Jeh, G., Widom, J.: Scaling personalized web search. In: WWW, pp. 271–279 (2003)
3. Jeh, G., Widom, J.: Simrank: a measure of structural-context similarity. In: KDD, pp. 538–543 (2002)
4. Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In: PVLDB, vol. 4(11), pp. 992–1003 (2011)