A Survey of Entity Similarity Measures on Heterogeneous Information Network

ZHU Zichen
Department of Computer Science
University of Hong Kong
Pokfulam Road, Hong Kong
zczhu2@cs.hku.hk

MA Chenhao
Department of Computer Science
University of Hong Kong
Pokfulam Road, Hong Kong
chma2@cs.hku.hk

ABSTRACT

According to recent studies, *heterogeneous information networks (HINs)* consisting of multiple types of entities and relations has shown its power in many disciplines, such as, computer science, social science, physics and so on. More and more researchers have noticed the importance of HIN analysis and many novel data mining tasks have been exploited in such networks, such as similarity search, clustering, and classification. Among those tasks, similarity measure on HINs, which is mainly to evaluate the similarity of entities, is the basis of many data mining tasks, such as clustering and classification. In this paper, we provide a survey of similarity measures on heterogeneous information networks. We will introduce basic concepts of heterogeneous information networks analysis, examine the recent developments on similarity measures and make a general evaluation of those similarity metrics.

1. INTRODUCTION

Recently, researchers use heterogeneous information networks (HINs) to model real world relationships in many applications, especially for real systems containing multi-typed interacting components. For instance, in bibliographic database, like DBLP[1] [5], papers are connected together via authors, venues and terms; and in Instagram[2], photos or videos are linked together via users, locations, hashtags and comments. Compared to widely-used homogeneous information networks [1, 2], which are extracted from real interacting systems by simply ignoring the heterogeneity of objects and links or only considering one type of relations among one type of objects, the heterogeneous information network can effectively fuse more information and contain rich and specific semantics in nodes and links [4].

Because of HIN's property of rich semantics and information, since the concept of heterogeneous information network and meta path proposed in 2009 [7] and 2011 [6], respectively, more and more researchers have noticed the importance of heterogeneouse information network analysis and many novel data mining tasks have been developed in such networks, such as similarity search [3, 6], clustering [8]. In other words, HIN analysis has become a hot topic rapidly in the fields of data mining, database and information retrieval, involving similarity measure, clustering, classification, link prediction, ranking, recommendation and information fusion on HINs [4].

Among those analysis tasks, similarity measure is the fundamental problem of network analysis, because most high level tasks, such as clustering and classification, need to evaluate the similarity of objects or relations. What's more, most of the state-of-the-art homogeneous network similarity methods, which generally assume the networks don't carry semantics, do not generalize well in HINs, due to HINs' rich semantics. Thus, some similarity metrics based on meta-path, which represents semantics in HINs, have been designed to evaluate the similarity between entities or relations in HINs, such as PathSim [6] and RelSim [9]. Maybe some figure of HIN and an example...

In this paper, we attempts to clearly introduce basic concepts in heterogeneous network analysis and make a conprehensive investigation on contemporary research developments of similarity metrics on HINs. Then, we tentatively provide a general evaluation over some of these state-of-the-art algorithms to give some reference on similarity measure choosing in high level network analysis tasks, such as clustering and classification.

The following part is organized as follows. Section 2 introduces the basic concepts and examples about HIN. Section 3 presents recent designed similarity measures on HIN. Experiments and evaluation are conducted in Section 4. Finally, Section 5 summarizes and concludes this paper.

[1] http://dblp.uni-trier.de
[2] http://instagram.com

COMP 9102 Surveys 2017, CS Dept, HKU

2. REFERENCES

[1] V. Leroy, B. B. Cambazoglu, and F. Bonchi. Cold start link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 393–402. ACM, 2010.

[2] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international*

conference on Knowledge discovery and data mining, pages 243–252. ACM, 2010.

[3] C. Shi, X. Kong, P. S. Yu, S. Xie, and B. Wu. Relevance search in heterogeneous networks. In *Proceedings of the 15th International Conference on Extending Database Technology*, pages 180–191. ACM, 2012.

[4] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):17–37, 2017.

[5] Y. Sun and J. Han. Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD Explorations Newsletter*, 14(2):20–28, 2013.

[6] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11):992–1003, 2011.

[7] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 565–576. ACM, 2009.

[8] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu. Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(3):11, 2013.

[9] C. Wang, Y. Sun, Y. Song, J. Han, Y. Song, L. Wang, and M. Zhang. Relsim: relation similarity search in schema-rich heterogeneous information networks. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 621–629. SIAM, 2016.