[APWeb-WAIM 2017 Tutorial Proposal] Meta Paths and Meta Structures: Analysing Large Heterogeneous Information Networks

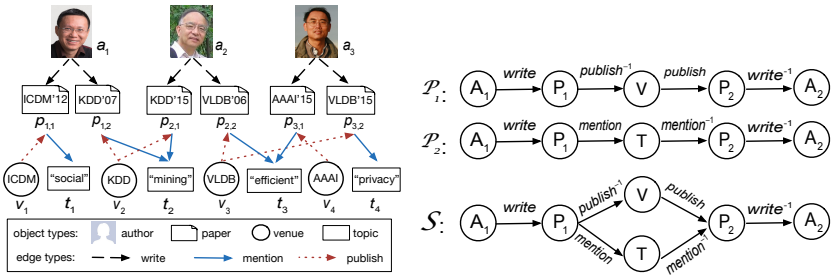
Reynold Cheng, Zhipeng Huang, Yudian Zheng, Jing Yan,
Ka Yu Wong, and Eddie Ng

University of Hong Kong, Pokfulam Road, Hong Kong,
{ckcheng,zphuang,ydzheng2,jyan}@cs.hku.hk,
kywong2@connect.hku.hk,ngheii@gmail.com,
WWW home page: http://www.cs.hku.hk/~ckcheng/

Abstract. A heterogeneous information network (HIN) is a graph model in which objects and edges are annotated with types. Large and complex databases, such as YAGO and DBLP, can be modeled as HINs. A fundamental problem in HINs is the computation of closeness, or relevance, between two HIN objects. Relevance measures, such as PCRW, PathSim, and HeteSim, can be used in various applications, including information retrieval, entity resolution, and product recommendation. These metrics are based on the use of meta-paths, essentially a sequence of node classes and edge types between two nodes in a HIN. In this tutorial, we will give a detailed review of meta-paths, as well as how they are used to define relevance. In a large and complex HIN, retrieving meta paths manually can be complex, expensive, and error-prone. Hence, we will explore systematic methods for finding meta paths. In particular, we will study a solution based on the Query-by-Example (QBE) paradigm, which allows us to discover meta-paths in an effective and efficient manner.
We further generalise the notion of meta path to "meta structure", which is a directed acyclic graph of object types with edge types connecting them. Meta structure, which is more expressive than the meta path, can describe complex relationship between two HIN objects (e.g., two papers in DBLP share the same authors and topics). We will discuss three relevance measures based on meta structure. Due to the computational complexity of these measures, we also study an algorithm with data structures proposed to support their evaluation. Finally, we will examine solutions for performing query recommendation based on meta-paths. We will also discuss future research directions.

1 Background

Heterogeneous information networks (HINs), such as DBLP [5], YAGO [8], and DBpedia [1], have recently received a lot of attention. These data sources, containing a vast number of inter-related facts, facilitate the discovery of interesting knowledge [4, 6, 7]. Figure 1(a) illustrates an HIN, which describes the relationship among entities of different types (e.g., author, paper, venue and topic). For example, `Jiawei Han` (a_2) has *written* a `VLDB` paper ($p_{2,2}$), which *mentions* the topic "`efficient`" (t_3).



(a) (b)

Fig. 1: HIN, Meta Paths, and Meta Structures.

Given two HIN objects a and b, the evaluation of their *relevance* is of fundamental importance. This quantifies the degree of closeness between a and b. In Figure 1(a), `Jian Pei` (a_1) and `Jiawei Han` (a_2) have a high relevance score, since they have both published papers with keyword "`mining`" in the same venue (`KDD`). Relevance finds its applications in information retrieval, recommendation, and clustering [9, 10]: a researcher can retrieve papers that have high relevance in terms of topics and venues in DBLP; in YAGO, relevance facilitates the extraction of actors who are close to a given director. As another example, in entity resolution applications, duplicated HIN object pairs having high relevance scores (e.g., two different objects in an HIN referring to the same real-world person) can be identified and removed from the HIN.

Relevance computation. In this tutorial, we will explore different ways of computing the relevance between two graph objects, for instance, neighborhood-based measures, such as *common neighbors* and *Jaccard's coefficient*; graph-theoretic measures based on random walks, such as *Personalized PageRank* and *SimRank*. These measures do not consider object and edge type information in an HIN. We will discuss the concept of *meta paths* [4, 9]. A meta path is a sequence of object types with edge types between them. Figure 1(b) illustrates a meta path $\mathcal{P}_1$, which states that two authors (A_1 and A_2) are related by their publications in the same venue (V). Another meta path $\mathcal{P}_2$ says that two authors have written papers containing the same topic (T). We will discuss several meta-path-based relevance measures, including *PathCount*, *PathSim*, and *Path Constrained Random Walk (PCRW)* [4, 9]. These measures have been shown to be better than those that do not consider object and edge type information.

We will further discuss *meta structures*, recently proposed in [3], to depict the relationship of two graph objects. This is essentially a directed acyclic graph of object and edge types. Figure 1(b) illustrates a meta structure $\mathcal{S}$, which depicts that two authors are relevant if they have published papers in the same venue, and have also mentioned the same topic. A meta path (e.g., $\mathcal{P}_1$ or $\mathcal{P}_2$) is a special case of a meta structure. However, a meta path fails to capture such complex relationship that can be conveniently expressed by a meta structure (e.g., $\mathcal{S}$). We will discuss how meta structures can be used to formulate three relevance definitions, as well as their efficient calculation.

Meta path discovery. There are often a huge number of meta paths between a pair of HIN objects. This can be very difficult, even for a domain expert, to identify the right meta paths. We will discuss a meta path discovery algorithm, recently proposed by [6],

where users provide example instances of source and target objects through a *Query-by-Example* paradigm, to derive meta paths automatically. We will demonstrate a HIN search engine prototype based on this algorithm.

Query recommendation. We will study the use of meta paths in query recommendation, where queries are suggested to web search users based on their previous query histories. As studied in [2], it is possible to use a knowledge base (a HIN) and its related meta-paths to perform effective query recommendation. The approach is especially useful to *long-tail queries* that rarely appear in query logs.

2 Proposed Schedule

The following is our proposed schedule of the 90-minute tutorial.

- **Introduction (15 minutes)**. We will discuss the basic model of HIN, and discuss applications based on it, such as search, relevance computation, query recommendation, and data integration (10 minutes). We will also introduce meta-paths, a fundamental HIN analysis tool, and give an overview of the tutorial (5 minutes).
- **Main contents (60 minutes)**. Next, we will introduce meta path, and how it facilitates the computation of various relevance measures (10 minutes). We then explain the process of discovering meta paths (15 minutes). We discuss a novel query recommendation framework based on meta paths (15 minutes). We will also present the meta structures, which is the latest development of meta paths (15 minutes). We will demonstrate a HIN search engine prototype based on meta paths (5 minutes).
- **Conclusions (15 minutes)**. We will conclude the tutorial and discuss future directions (5 minutes). The rest of the time will be dedicated to Q&A (10 minutes).

3 Intended Audience

The tutorial is designed for researchers interested in latest development in the field of HINs, especially regarding meta-paths for novel applications. The HIN search demonstration will be give insight to software practitioners for developing recommendation facilities for HINs.

4 Biography of Presenters

Reynold Cheng is an Associate Professor of the Department of Computer Science in the University of Hong Kong. He obtained his PhD from Department of Computer Science of Purdue University in 2005. He was granted an Outstanding Young Researcher Award 2011-12 by HKU. He was the recipient of the 2010 Research Output Prize in the Department of Computer Science of HKU. He also received the U21 Fellowship in 2011. He received the Performance Reward in years 2006 and 2007 awarded by the Hong Kong Polytechnic University. He is a member of the IEEE, the ACM, and ACM SIGMOD. He is an editorial board member of TKDE, DAPD and IS, and was a guest editor for TKDE, DAPD, and Geoinformatica. He is an area chair of ICDE 2017, senior

PC member of BigData 2017 and DASFAA 2015, PC co-chair of APWeb 2015, area chair for CIKM 2014, and workshop co-chair of ICDE 2014. He received an Outstanding Service Award in CIKM 2009. He has served as PC members and reviewer for top conferences and journals.

Zhipeng Huang is a 2nd year Ph.D. in the CS department of HKU, supervised by Prof. Nikos Mamoulis and Dr. Reynold Cheng. He received his bachelor degree from EECS department of PKU in 2015. His research interests cover data mining, data management and data cleaning.

Yudian Zheng is a 4th year Ph.D. in the CS department of HKU, supervised by Dr. Reynold Cheng. Yudian's research interests cover crowdsourcing, data management and data cleaning. He has published full research papers in well-established database and data mining conferences/journals, including SIGMOD, VLDB, KDD, WWW, ICDE, and TKDE. He has also taken internships in Microsoft Research and Google Research.

Jing Yan is a 1st year MPhil student supervised by Dr. Reynold Cheng in the CS department of HKU. His research interests include data management and data mining, with emphasis on knowledge graphs and data cleaning.

Ka Yu Wong is currently a MSc student of the CS department of HKU.

Eddie Ng is currently a MSc student of the CS department of HKU.

Acknowledgements

Reynold Cheng, Zhipeng Huang, Yudian Zheng, and Jing Yan were supported by the Research Grants Council of Hong Kong (RGC Projects HKU 17229116 and 17205115) and the University of Hong Kong (Projects 102009508 and 104004129).

References

1. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.
2. Z. Huang, B. Cautis, R. Cheng, and Y. Zheng. Kb-enabled query recommendation for long-tail queries. In *CIKM*, pages 2107–2112, 2016.
3. Z. Huang, Y. Zheng, R. Cheng, Y. Sun, N. Mamoulis, and X. Li. Meta structure: Computing relevance in large heterogeneous information networks. In *KDD*, 2016.
4. N. Lao and W. W. Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine learning*, 81(1):53–67, 2010.
5. M. Ley. Dblp computer science bibliography. 2005.
6. C. Meng, R. Cheng, S. Maniu, P. Senellart, and W. Zhang. Discovering meta-paths in large heterogeneous information networks. In *WWW*, pages 754–764, 2015.
7. D. Mottin, M. Lissandrini, Y. Velegrakis, and T. Palpanas. Exemplar queries: Give me an example of what you need. *PVLDB*, 7(5):365–376, 2014.
8. F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *WWW*, pages 697–706, 2007.
9. Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *PVLDB*, pages 992–1003, 2011.
10. X. Yu, X. Ren, Y. Sun, B. Sturt, U. Khandelwal, Q. Gu, B. Norick, and J. Han. Recommendation in heterogeneous information networks with implicit user feedback. In *RecSys*, pages 347–350, 2013.