COMP4801- Final Year Project

Project Plan

Project Title: Discovering and Querying Meta-Graphs in Large Heterogeneous Information Networks

Student: Fung Yuet (UID: -)

Supervisor: Dr. Reynold C.K. Cheng

Table of Content

Introduction

Data have been increasingly available as the use of the Web and social media by people has been surging. This copious amount of rich information enables the study of knowledge discovery in data aimed for better understanding of the human behaviour. For instance, researchers mine patterns from the training data for the use of trend projection; researchers investigate the topic of relevance between objects to enhance the similarity search. In recent decades, mining in information networks has aroused growing attentions from researchers because many datasets can be organized into a graph whose complex structural characteristics allows one to dig out promising knowledge. An information network is a graph $G = (V, E)$, where V is the set of nodes (i.e. *objects*) and E is the set of edges (i.e. *relationships*). An example is depicted in Figure 1(a). For each node, it has one or more associated node classes; for each edge, it has one or more associated edge types. Thus, there are two more important functions: $\chi: V \rightarrow \mathcal{L}$; $\psi: E \rightarrow \mathcal{R}$. The first function χ is the node class matching function where each object $v \in V$ matches to one or more node classes $\chi(v) \in \mathcal{L}$. Likewise, the second function ψ is the edge type matching function where each edge $e \in E$ belongs to one or more edge types $\psi(e) \in \mathcal{R}$. Accordingly, there are two kinds of information network:

if $|\mathcal{L}| = 1$ and $|\mathcal{R}| = 1$, it is a homogeneous information network

if $|\mathcal{L}| > 1$ or $|\mathcal{R}| > 1$, it is a heterogeneous information network (*aka HIN*) [14]

Homogeneous information network mining has been studied since the last few decades during which researchers have been developing methods for analysing homogeneous information networks on the task of clustering, ranking and link prediction [14]. It although seems feasible to extend some of these techniques to handle the study of HINs, most of them cannot be directly applied to the problem. It is because the schema of an HIN is far more complicated than the one in a homogeneous information network and this enables an HIN to express richer information. In addition, node classes and edge types are different across objects and relations. Consequently, considering them as identical as those in the case of homogeneous information network loses the semantic meaning and possibly the valuable information one would have mined from the network. Therefore, researchers should develop a new set of methodologies and principles in studying HINs.

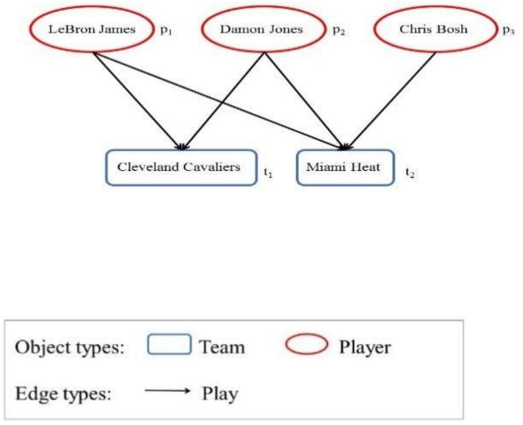

Figure 1(a) HIN

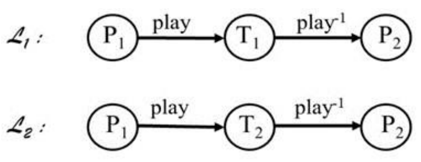

Figure 1(b) Meta paths

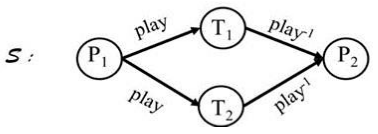

Figure 1(c) Metagraph

One interesting research topic in HINs analysis is relevance and similarity search. Relevance of two objects defines how similar they are or how tightly they are being related. These research results provoke the analysis of similarity search, classification, clustering and link prediction in HINs [14]. While many works have been done on relevance study, e.g. *Personalized PageRank* [5], *Jaccard's coefficient* [11] and *SimRank* [6], these measures neither consider the different semantic meanings of node classes nor that of edge types. Regarding this issue, Sun. et al [17] proposed the concept of *meta path* and a family of *meta path-based similarity measures.* A meta path is a path comprising of node classes and edge types instead of the actual objects and relations. Node classes and edge types are called meta data in network analysis and thus a path of meta data is recognized as a meta path. Two examples are illustrated in Figure 1(b).

Although meta paths have been proved to be useful in many applications, it can only capture simple relationships between the source node and the target one. More importantly, it simplifies all the relations as a single path relation and overlooks the semantic different between a combination of meta paths and a structural relation. Huang et al. [4] subsequently proposed a concept of *meta structure* aiming to provide a data structure depicting complex relations; Fang et al. [2] at the mean time suggested *metagraph*, being a generalisation of meta path and meta structure, to represent various semantic relations. An example of metagraph is shown in Figure 1(c). Meta-graphs (*i.e. meta path, meta structures and metagraph*) have then been proved to be beneficial to product recommendation, community detection, classification and clustering and link prediction [1-4,7-8,10].

Despite the substantial studies of meta-graphs application, most of the works assumed these structures are given by experts or are discovered using enumeration or Breadth First Search. Sun et al. [15] assumed the meta paths defining the co-author relationship is given by experts in the study of applying meta paths in prediction and recommendation. Yu et al. [18] and Li et al. [10] both presumed the meta paths will be given in the analysis of recommendation and clustering. Although some researchers attempted to use Breadth First Search to discover the meta path [8,16], they require the use to input the maximum meta path length restricting the search space. Similarly, researchers expected *metagraph* are given by domain experts [19] or by enumeration [7] for the study of its applications. Since meta-graphs are difficult to define for a complex schema and it is a labour-demanding task [12], while Breadth First Search and enumeration are not efficient for large HINs, it is obvious that researchers should develop an efficient approach to discover these meta-graphs automatically in large heterogeneous information networks.

Some researchers have studied this problem in the last few years attempting to automatically discover and locate these illustrative data structure in large HINs. Regarding the discovery of meta paths, there are in general two approaches: *example-based training* and *adapted sequential pattern mining*. Example-based training means the users need to first provide positive example pairs. The algorithm framework will train the model based on these pairs and will transverse the HIN to discover the meta paths highly explaining the relationship of the given pairs. There are many proposed algorithms using this framework, e.g. FSPG [12], AMPG [1] and SMPG [20], though they differ subtly in some areas. For example, the definitions of the priority score for heuristics pruning and the method of discarding unimportant meta paths. Additionally, Shi et al. [13] proposed a method adapted from well-known knowledge discovery techniques – sequential pattern mining, aiming to simulate mining interesting meta paths as sequential pattern mining. “Generate-and-Discard” suggested by Shi [13] targets to generate meta paths linking the two target nodes by considering the linkage between the siblings of the two target nodes.

Although these two approaches can efficaciously discover meta paths with the aid of efficient data structure, they are neither scalable nor comprehensive in handling the metagraphs discovery. Since example-pair based approach transverses the HIN to locate possible candidates, it might not be scalable in discovering meta-graphs with complex schema since during each branching step, there are many more possible candidates to consider. Even though the algorithmic framework considers further searching on one candidate, the intermediate calculations and maintenance would cost a lot. This method may be scalable for some special cases of a meta-graph. For example, a symmetric meta-graph or a meta-graph with simple structure. It is not comprehensive in handling the generalisation version. Likewise, the “generate-and-discard” may not scale well as the number of possible candidates goes exponential when it is being applied to meta-graphs.

Regarding the discovery of meta-graphs, Fang et al. [2] proposed a heuristic approach to mine meta-graphs from HINs. They proposed to use a set of seed candidate of meta-graphs, assuming that two structurally similar meta-graphs are functionally similar. By selecting candidates with largest candidate heuristic score means that the chosen one is structurally and functionally similar to any seed meta-graphs. However, it is unclear whether functional similarity and structural similarity are highly correlated. Moreover, using a set of seed meta-graphs would limit the diversity of candidates. It thus needs more justifications. In short, researchers should develop a more unified and efficient framework in handling automatic discovery of meta-graphs for large heterogeneous information networks.

Problem Statements & Objectives

The goal of this work is to develop a systematic and methodical algorithmic framework allowing efficient discovery of meta-graphs in large heterogeneous information networks. In this work, the method proposed in [12] will be adapted with modifications for optimization in the discovery. Below are the definitions of models used in this study and the specific objectives of this work.

Problem Definitions

DEFINITION 1 (HETEROGENEOUS INFORMATION NETWORK).

A heterogeneous information network is a graph G = (V, E), where V is the set of nodes (i.e. objects) and E is the set of edges (i.e. relationships). There are two more important functions: $\chi: V \rightarrow \mathcal{L}$; $\psi: E \rightarrow \mathcal{R}$. *The first function* χ *is the node class matching function where each object* $v \in V$ *matches to one or more node classes* $\chi(v) \in \mathcal{L}$. *Likewise, the second function* ψ *is the edge type matching function where each edge* $e \in E$ *belongs to one or more edge types* $\psi(e) \in \mathcal{R}$. *Note that* $|\mathcal{L}| > 1$ or $|\mathcal{R}| > 1$.

DEFINITION 2 (META-GRAPHS).

Given an HIN G = (V, E, $\mathcal{L}$*,* $\mathcal{R}$*) with* χ and ψ, *a meta-graph is a graph* $\bar{G} = (\bar{V}, \bar{E})$ *where* $\bar{V} \in \mathcal{L}$ *and* $\bar{E} \in \mathcal{R}$.

PROBLEM 1 (RELEVANT META-GRAPHS).

Given an HIN $G = (V, E, \mathcal{L}, \mathcal{R})$ *with* χ and ψ, *together with a set of n example pairs* $S_{ep} = \{(s_i,t_i) \mid i \in [1, n]\}$ *and a similarity function* φ, *discover a set of m meta-graphs* $S_{mg} = \{g_i \mid i \in [1, m]\}$ *that capture the characteristics of each pair in* S_{ep}.

Objectives

- To design optimization techniques which will be adapted in the approach proposed in [12]
- To implement the proposed framework as an application of "Query-by-example"

Methodology

In this study, we will adapt an existing algorithm framework proposed by Meng et al. [12] but will extend it to the discovery of meta-graphs with several enhancements and optimisations being presented. Meng et al. [12] suggested a two-phase framework in finding meta paths in large HINs. First, the authors illustrated how to use a modified Least-Angle Regression model to progressively learn relevant link-only meta-paths. The model will gradually select the meta-paths with the largest correlation values. Second, to generate link-only meta-paths, the authors designed a novel data structure named "GreedyTree". With the aid of the priority score, the algorithm can then generate link-only meta-paths. Finally, by considering the Least Common Ancestor of the unknown classes in the class hierarchy and TF-IDF weighting, a set of relevant meta-paths would have been generated.

However, this approach may not scale well under the discussion of meta-graphs as the number of candidates rises exponentially. Our proposed approach is to implement heuristic functions and efficient data structures enabling efficacious discovery of meta-graphs in HINs in the average case.

Schedule

Sept 2017 – Jan 2018	Development of the algorithmic framework
Feb 2018	Empirical Studies
Mar 2018 – Apr 2018	Refinements on the framework
May 2018	Query-by-Example Application

Reference

[1] X. Cao, Y. Zheng, C. Shi, J. Li and B. Wu, "Meta-path-based link prediction in schema-rich heterogeneous information network", *International Journal of Data Science and Analytics*, vol. 3, no. 4, pp. 285-296, 2017.

[2] Y. Fang, W. Lin, V. Zheng, M. Wu, K. Chang and X. Li, “Semantic proximity search on graphs with metagraph-based learning”, in *ICDE,* 2016.

[3] Z. Huang, B. Cautis, R. Cheng and Y. Zheng, “KB-Enabled Query Recommendation for Long-Tail Queries”, in *CIKM,* 2016.

[4] 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.

[5] G. Jeh and J. Widom, “Scaling Personalized Web Search”, in *WWW,* 2003.

[6] G. Jeh and J. Widom, “SimRank: A Measure of Structural-Context Similarity”, in *KDD,* 2002.

[7] H. Jiang, Y. Song, C. Wang, M. Zhang and Y. Sun, “Semi-supervised Learning over Heterogeneous Information Networks by Ensemble of Meta-graph Guided Random Walks”, *IJCAI,* 2017.

[8] X. Kong, B. Cao, P. Yu, Y. Ding and D. Wild, “Meta Path-Based Collective Classification in Heterogeneous Information Networks”, in *CIKM,* 2012.

[9] N. Lao and W. Cohen, “Relational Retrieval using a combination of path-constrained random walks”, *Machine Learning,* 2010.

[10] X. Li, Y. Wu, M. Ester, B. Kao, X. Wang and Y. Zheng, “Semi-supervised Clustering in Attributed Heterogeneous Information Networks”, in *WWW,* 2017.

[11] D. Nowell and J. Kleinberg, “The link-prediction problem for social networks”, *J. Assoc. Inf. Sci. Technol.,* 58(7), 2007.

[12] C. Meng, R. Cheng, S. Maniu, P. Senellart, and W. Zhang, “Discovering meta-paths in large heterogeneous information networks,” in *WWW,* 2015, pp. 754–764.

[13] B. Shi and T. Weninger, “Mining interesting meta-paths from complex heterogeneous information networks,” in *ICDM-MODAT,* 2014.

[14] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, “Co-author relationship prediction in heterogeneous bibliographic networks,” in *ASONAM,* 2011, pp. 121–128.

[15] Y. Sun and J. Han, *Mining Heterogeneous Information Networks: Principles and Methodologies*. Morgan & Claypool Publishers, 2012.

[16] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla, “When will it happen? Relationship prediction in heterogeneous information networks,” in *WSDM,* 2012, pp. 663–672.

[17] Y. Sun, J. Han, X. Yan, P. Yu, and T. Wu, “Pathsim: Meta path-based top-k similarity search in heterogeneous information networks,” in *VLDB,* 2011, pp. 992–1003.

[18] 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,* 2013, pp. 347–350.

[19] H. Zhao, Q. Yao, J. Li, Y. Song and D. Lee, “Meta-Graph Based Recommendation Fusion over Heterogeneous Information Networks”, in *KDD,* 2017, pp. 634-655.

[20] Y. Zheng, C. Shi, X. Cao, X. Li and B. Wu, “Entity Set Expansion with Meta Path in Knowledge Graph”, in *PAKDD,* 201, pp. 317-329.