Semantic data mining using RDF hypergraphs

Wei Yu (2731**) Hanxiao Liu (2731****)**

Submission date: 2016-05-13

Southeast University-Monash University Joint Graduate School, Suzhou, China

Abstract

This paper focuses on techniques and applications of semantic data mining using RDF hypergraphs. First, fundamental techniques about the graph-based semantic data mining and the issues of large graphs are introduced. Second, a concept of RDF hypergraphs is proposed in biomedical semantic data mining, which aims at discovering semantic associations and detecting potential errors about semantic associations between drugs and diseases. Third, we also review the application of mining semantic connections in the social network. Finally, we advance an assumption to apply techniques of mining semantic associations in social network to biomedicine area.

Keywords: Semantic data mining, RDF hypergraphs, Semantic association

1. Introduction

The emergence of World Wide Web (WWW), proposed by Tim Berners-Lee, has largely changed people's life in the worldwide. There are billions of resources stored in the web which could have been more efficiently utilized. However, the current situation is that they are more like isolated islands rather than interconnecting with each other, because they are only readable by human being and it is impractical to handle with them manually. Therefore, the Resource Description Framework (RDF) is proposed to describe resources in the web in the form of subject-predicate-object expressions [1]. Compared with conventional World Wide Web, it is machine-readable and through which machine can store and exchange information. A RDF document is usually looked as a graph and RDF graphs have several representations such as labeled directed graphs (LDG) [1]. It is easy for LDG to describe simple examples. However, RDF graphs cannot depict complex cases [2]. Due to limitations of fundamental RDF graphs, Hypergraphs for RDF are introduced as a representation of web resources. The concept of hypergraph is generalized from graphs, hypergraphs can be efficiently employed in many fields like social networks and database theory [3]. Hence, the combination of hypergraphs and semantic web, which is based on RDF, can be widely used to improve the efficiency and accuracy of semantic data mining. The semantic data mining is similar to standard data mining, but it is usually employed with graphs [3]. This thesis mainly reviews several papers of semantic data mining in the field of social network and biomedicine. Besides, we introduce algorithms proposed in each paper and analyze advantages and limitations of them. The main contribution of this thesis is that it provides an overview of current techniques and application of the semantic data mining using RDF hypergraphs. In the

following parts, section 2 introduces the scope and method of the thesis. Section 3 reviews five papers which applied in the field of semantic data mining. In section 4, we summarize the whole thesis.

2. Scope and Method

In this thesis, all literatures are selected from Monash library and Google Scholar, the criterion of the selection is whether the literature has strong correlation with our research topic or not. We mainly focus on current techniques and applications which are applied in the field of biomedicine of semantic data mining. we also introduce applications of semantic data mining in the area of social network, because we think that algorithms which employed in social network can be similarly applied and have effectively implication on the research work in biomedicine. Before the selection of literatures, we select semantic data mining using RDF hypergraphs as our research topic. Then, we focus our attention on two dimensions: semantic data mining and hypergraphs. First, we find a paper which introduces hypergraphs for RDF and has mentioned limitations of fundamental RDF graphs. Second, we review a paper which summarizes fundamental algorithms of directed hypergraphs [3]. Then, we begin to search the concrete application in the semantic data mining of hypergraphs for RDF. We notice that the most application of semantic data mining is lying in biomedicine and social network. We therefore look through 5 papers which introduce their current studies in mining biomedical ontologies and data in the biomedicine area [12] and mining multi-layer networks for connectivity patterns in the social network area [4].

3. Review

3.1 Fundamental techniques of semantic data mining

To discover semantic associations between objects, Liu [10] introduces a graph-based approach. In this article, he first points out that cooperation between domain knowledge and data mining is still at a primary stage. To solve this problem, he develops a framework called semantic data mining. This framework has three contributions. First, he describes how to use RDF model and its graph representations to describe domain knowledge and data. Second, he proposes a method which is based on metaheuristic optimization to deal with heterogeneous data. Third, he shows the application of semantic data mining framework on discovering the semantic associations between objects. Finally, the author points out the limitations of the semantic data mining framework and introduces the future work. Three aspects are mentioned as following. First, the author addresses that current semantic data mining framework lacks automatic and robust semantic annotations [10]. As a result, the author puts forward a primary idea of learning based semantic search method. Second, the author points out that the weight of relationships between ontologies cannot be automatically determined by system [10]. It usually depends on experts' empiricism. Third, scalability issues are big problems. In graph-based semantic mining framework, practical problems are often in large scales. It is hard to efficiently process the graphs if they represent the ontologies and data in practical issues with large scale. This paper proposes RDF hypergraphs to represent ontologies and data in semantic data mining which improve accuracy and scalability of the result. However, it does not make experiments to

compare the graph-based approach with other approaches which makes it hard to determine which approach performs better in different applications.

Recently, information volume is increasing sharply and relationships between the information become more and more complex. Graphs which are applied to describe the information and relationships between the information are larger and larger. Mining large graphs currently is important in various application domains. The vast graph data is at risk going beyond commodity servers load capacity. To tackle this problem efficiently, we can partition big graphs and process the partitions of big graphs in distributed servers. Yang et al. [11] provide a method called Self Evolving Distributed Graph Management Environment (Sedge) to handle with large graphs. Sedge employs two-level partition management architecture which includes complementary primary partitions and dynamic secondary partitions [11]. It aims to cut down inner communications between machines, make systems respond in time and improve the throughput of systems [11]. They also mention that the Sedge can be utilized to adjust partitions to server queries instead of proposing a new graph partition algorithm. The Sedge method, based on Pregel which is distributed and fault-tolerant, addresses the graph partition management. It adds overlapping partitions function compared with Pregel. This method also provides a solution to the problem which is described in previous article. This paper provides a precise evaluation on performance and concrete realization of the experiment.

3.2 Semantic data mining in biomedicine.

Semantic data mining has been extensively carried out in various research areas, especially in biomedicine area where domain knowledge evolves rapidly [12]. Liu et al. [12] put forward an idea by using the RDF hypergraphs to mine biomedical ontologies and data. The mining aims at solving two problems: Discovering semantic associations using ontologies and detecting potential errors about semantic associations with ontologies and data. There are four steps in the process of discovering semantic associations between objects in biomedical area. First, using RDF hypergraphs to express biomedical ontologies and data. Second, transforming RDF hypergraphs into bipartite graphs. Third, utilizing matrixes to represent bipartite graphs and doing form transformation on matrixes. Fourth, applying random walk with restart algorithm to calculate the degree of semantic associations. Matrixes in the above are parameters in the random walk with restart algorithm. Then the paper uses the semantic associations' degree and the existing data to detect misinformation about semantic associations in ontologies. RDF triple has a graph nature [12], and a graph with a set of RDF triples is called directed labeled graph (DG). The paper adopts RDF hypergraphs rather than simple DG to represent biomedical ontologies and data, and one hypergraph edge can connect more than two vertices so that hypergraphs can represent more information. Moreover, RDF hypergraphs can eliminate disadvantages of DG. The most explicit disadvantage of DG is inconsistent representations. For example, arcs usually represent properties, but in meta-statement about properties, the properties must be represented as nodes [12]. The deficiency of this article is that authors do not put the method in practical situation. In practice, the data will be large enough, we can consider using Sedge which we

mention in 3.1 to deal with large graphs. In addition, authors do not compare the random walk with restart algorithm with other algorithms. They also have not explained why they use the random walk with restart algorithm.

3.3 Semantic data mining in social networks

Social network, as an unexploited treasure, has become an increasingly important research area. In this thesis, we will introduce two relevant papers.

In the first paper, Nettleton does a survey about data mining of social network represented as graphs [4]. He divides the survey into two main parts: the first part of survey includes literatures which provide the basis and background for the field of data mining of social network. The second part of survey encompasses hot topics which are prevalent within the research area. Throughout the whole paper, the literatures which focus on basis and background of the field have four main themes: graph theory, social network, online social networks and graph mining [4]. The graph theory summarizes key abstract concepts of graphs which construct as a basis of data mining. Online social network is helpful to create or define social relations among people. Graph mining, as an important part of the first part of the survey, also covers several key themes. Within key themes, algorithms like VF2 [5] for isomorphic matching in the data processing and Louvain method [6] for the community detection in the online social network communities are introduced to process the large data volume which has widely use in the online social network applications. At last, Nettleton mentions that online social network usually lacks complete picture of what users are doing, which will influence the accuracy of results of mining from the log datasets.

In today's social network, there are multiple sources of connectivity information between users. Oselio et al. [7] think that standard graph analyses are not qualified to represent social network relationships and multi-layer graphs are therefore employed to render the complex social network connectivity information. Due to that there is no advanced techniques to do semantic analysis in multi-layer graphs, they raise latent variable models and methods to mine connectivity patterns in multi-layer graphs. They mention that there are two type of sources which are included in the connectivity information: relational information and behavioral information. Compared with relational information, behavioral information is usually inferred from user's individual preferences or usage statistics which are not directly connect users. In the paper, they put forward a generative hierarchical latent-variable model to infer, cluster and detect semantic relationships in multi-layer graphs of social network based on techniques of Bayesian Model Averaging [8]. As a result, they can get the posterior probability of latent variables from the multi-layer network. Finally, they make a simulation which combined with the dynamic stochastic block model [9] and demonstrate that the model can explore complex connections within two layers of network. The methods proposed in this paper can effectively mine latent connections in the multi-layer social networks and it is scalable to explore other methods to infer multi-layer networks. However, the model only provides two-layer network to go on simulation in the experiment which is not persuadable to confirm that the model works in the three-layer or more layer's social network.

4. Conclusion

In this review, we focus on the application of RDF hypergraphs in semantic data mining. we review five papers about semantic data mining in the field of biomedicine and online social network. First we introduce fundamental techniques about graph-based semantic data mining and techniques of dealing with large graph. Then we review an article which introduces the function of semantic data mining in biomedicine area and analyze shortcomings of the method it develops. Afterwards, we review two papers in the field of social network. The first paper does a survey about semantic data mining of social networks with the presentation of graphs. In the second paper, based on techniques of Bayesian Model Averaging, authors raise latent variable models and methods to mine connectivity patterns in multi-layer graphs to infer, cluster and detect semantic relationships in the multi-layer graphs of social. The purpose of reviewing papers in the social network of semantic data mining is to apply the model proposed in this field to solve the similar problems in biomedicine.

References

[1] Morales, A., & Serodio, M. (2007). A directed hypergraph model for RDF. Proc. of Knowledge Web PhD Symposium.

[2] Hayes, J., & Gutierrez, C. (2004). Bipartite graphs as intermediate model for RDF. The Semantic Web–ISWC 2004, pp. 47-61.

[3] Ausiello, G., & Laura, L. (2016). Directed hypergraphs: Introduction and fundamental algorithms—A survey. Theoretical Computer Science.

[4] Nettleton, D. F. (2013). Data mining of social networks represented as graphs. Computer Science Review, 7, pp. 1-34.

[5] Cordella, L.P., Foggia, P., Sansone, C., & Vento, M. (2001). An Improved Algorithm for Matching Large Graphs, Proc. 3rd IAPRTC-15 International Workshop on Graph based Representations, Cuen, Italy, 2001, pp. 149–159.

[6] Blondel, V.D., Guillaume, J.L., Lambiotte, R., & Lefebure, E. (2008). Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment, vol. 10, 1000.

[7] Multi-Layer Graph Analysis for Dynamic Social Networks.

[8] Raftery, A. (1995). “Bayesian model selection in social research,” Sociol. Methodol., vol. 25, pp. 111–164.

[9] Xu K.S. & A. O. H. III, (2013). “Dynamic stochastic blockmodels: Statistical models for time-evolving networks,” CoRR, vol. abs/1304.5974.

[10] Liu, H. (2012). A graph-based approach for semantic data mining.

[11] Yang, S., Yan, X., Zong, B., & Khan, A. (2012). Towards effective partition management for large graphs. Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 517-528.

[12] Liu, H., Dou, D., Jin, R., LePendu, P., & Shah, N. (2013). Mining biomedical ontologies and data using RDF hypergraphs. Proceedings of the 2013 12th IEEE International Conference on Machine Learning and Applications (ICMLA), vol. 1, pp. 141-146.