Semantic data mining using RDF hypergraphs

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

Submission date: 2016-05-30

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

Abstract

In recent year, the amount of data which is generated by users is increasing with a high speed. It is critical for researchers to design efficient methods to extract semantic associations from the generated big data which is usually scattered, redundant, and mutually complementary [1]. This paper introduces some relevant techniques to mine sematic associations in the field of biomedicine. we firstly propose to utilize the resource description framework (RDF) hypergraph to mine biomedical ontologies and data given that it can represent connections between more than two vertices which can display more information compare with traditional RDF graphs. Then, we introduce a random walk with restart (RWR) algorithm and analyze its limitations the relevance score computation between two objects is not scalable in large graphs. Finally, a partition management mechanism and a fast random walk with restart (FRWR) algorithm are introduced to overcome the scalability problems of traditional RWR algorithm. Moreover, we also illustrate the novelty of this paper.

Keywords: Semantic associations, RDF Hypergraphs, FRWR, partition management mechanism

1. Introduction

Semantic data mining has been extensively harnessed in various areas, especially in biomedical area. In this area, it is significant to capture new associations and detect wrong relationships between drugs and diseases which can improve people's healthy level and prolong lifespan of people. Traditional graphs and RWR algorithms are widely utilized to perform above-mentioned work. However, there are three main limitations within current solutions. First, traditional graphs cannot be applied to represent co-occurrence relationships when the amount of objects is beyond two [2]. Second, Errors which exist in both ontologies and data are imperative to be tackled. Third, As the size of practical problems increases, RWR algorithm is not scalable for large graphs and requires more response time [3], algorithms aiming at searching and mining large graphs therefore need to be designed. Yang et al. propose to use effective partition management for large graph and Tong et al. put forward a FRWR algorithm to mine semantic associations in large graphs [3][4]. Contributions of this paper are as follows: we firstly employ hypergraphs to do semantic data mining, which is rarely utilized in biomedicine field. Then, we introduce a partition management and an improved RWR (FRWR) algorithm to compute the related score between

vertices in a graph while problems of large graphs are currently not taken into account in traditional RWR algorithms.

2. Objectives

Our research aims to develop a method which can discover semantic associations and detect potential errors between drugs and diseases in practical situations where the data volume is big. Moreover, we hope our method can improve the speed of the algorithm. Through reviewing some articles, we find that most of them adopt RWR algorithm. Tong et al. point out that straightforward application of RWR algorithm is not suitable for large graphs [3]. Therefore, in our research, a partition hypergraphs management mechanism and a fast solution (FRWR) to handle with problems of large graphs are introduced and the fast solution is reflected on two major dimensions, they are linear correlations and block-wise, community-like structure.

3. Methodology

First, we employ a RDF hypergraph to represent both biomedical ontologies and data. The definition of hypergraph is as follows: "a hypergraph $\mathrm{HG} = (\mathrm{V}, \mathrm{E})$, is a pair in which V is the vertex set and E is the hyperedge set where each $\mathrm{e} \in \mathrm{E}$ is a subset of V "[2]. A hypergraph is a generalization of a traditional graph. The hyperedge in hypergraphs can connect more than two vertices while an edge can only connect two vertices in a traditional graph. Therefore, hypergraphs can express co-occurrence relationships directly among more than two nodes. Any collection of ontologies and data can be represented as RDF graphs and any RDF graph also can be represented by a hypergraph. An RDF triple correspond to a hyperedge, the subject, predicate and object in an RDF triple denote vertices of the hyperedge.

Second, we partition the hypergraphs into several parts to make them have minimum connections. The commonly used methods are Pregel [5] and SPAR [6]. However, they have a bad performance when performing a random walk starting at a vertex. In our research, we adopt Self Evolving Distributed Graph Management Environment (Sedge) method which is proposed by Yang et al. to do partition work [4]. In Sedge, Yang et al. develop three main techniques. The three techniques are complementary partitioning, partition replication and dynamic partitioning. Complementary partitioning is used to discover different partition schemas. Partition replication is used to replicate the same partitions in different machines to share the workload on these partitions. Dynamic partitioning aims at constructing partitions to serve cross-partition queries locally. The Sedge can balance the workload on different machines and improve graph query response time and throughput.

Third, we need to translate the RDF hypergraph partitions into bipartite graphs. There are only two steps during the transforming process. First, divide the vertex set V and hyperedge set E into two parts. Second, connect $v_i \in \mathrm{V}$ and $e_i \in E$ if v_i is contained in e_i.

Finally, FRWR algorithm, instead of the FRWR algorithm, are employed in the paper to calculate the relevance score on bipartite graphs, which usually represent the semantic association degree between two objects. If the relevance score between two objects is over a given number, they are believed to have semantic association. At the same time, semantic associations are also exploited to detect potential errors between diseases and drugs in

biomedical area. When computing relevance score of two objects, it requires computing matrix inversion. Compared with traditional RWR algorithms, the FRWR algorithm improves two solutions of RWR algorithms when computing matrix inversion of large graphs. The first solution of the RWR algorithm is “precompute method” which is limited in space while it has a quick query time. The second solution is “onthefly method” of which the query time is slow [3]. In the FRWR algorithm, two properties of real graphs like linear correlations across rows and columns of the adjacency matrix and the block-wise, community-like structure are harnessed by authors, which can effectively balance the response quality and space cost.

4. Novelty

The novelty of this research is that we combine RDF representations and graph partition together to discover semantic associations between drugs and diseases in the biomedicine area. An improved RWR algorithm called the FRWR algorithm is also introduced to overcome the scalability problem of the RWR algorithm in large graphs. After reviewing a series of articles, we find that few articles adopt hypergraphs for semantic data mining. A hypergraph representation can display more information and the method of expressing the information is more directly than the traditional graph representation. Therefore, the hypergraph representation is adopted in our research. However, there are still some limitations existed in RDF and some solutions are also introduced in this paper. First, with the increment of ontologies and data, traditional RWR algorithms are not applicable to extract semantic associations in large graphs, which usually result in slow response time. Therefore, we introduce a partition management mechanism to partition large hypergraphs into several parts and distribute the partitions in different machines which can improve the efficiency of semantic data mining. Besides, a FRWR algorithm is introduced to overcome the limitation of traditional RWR combining with the partition management mechanism

5. Conclusion

In summary, this paper mainly focuses on the semantic data mining in the biomedicine area using RDF hypergraphs, which has positive effect in raising healthy level of human. We firstly explore the meaning of adopting hypergraphs in biomedicine area compared with traditional RDF graphs. Then, scalability problems of the RWR algorithm in large graphs are analyzed. Finally, we introduce a partition management mechanism and an improved RWR (FRWR) algorithm to compute the relevance score between vertices in a graph. This algorithm, exploiting two properties of real graphs, can solve problems of RWR algorithm by balancing response quality and space risk when computing the matrix inversion.

References

[1] Hsu, P. L., Hsieh, H. S., Liang, J. H., & Chen, Y. S. (2015). Mining various semantic relationships from unstructured user-generated web data. *Web Semantics: Science, Services and Agents on the World Wide Web*, *31*, 27-38.

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

[3] Tong, H., Faloutsos, C., & Pan, J. Y. (2006). Fast random walk with restart and its applications. *Proceedings of the International Conference on Data Mining,* pp. 613-622.

[4] 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.

[5] Malewicz, G., Austern, M. H., Bik, A. J., Dehnert, J. C., Horn, I., Leiser, N., & Czajkowski, G. (2010). Pregel: a system for large-scale graph processing. *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pp. 135-146.

[6] Pujol, J. M., Erramilli, V., Siganos, G., Yang, X., Laoutaris, N., Chhabra, P., & Rodriguez, P. (2011). The little engine (s) that could: scaling online social networks. *ACM SIGCOMM Computer Communication Review*, *41*(4), pp. 375-386.