Study of Hadoop-Based Semantic Big Data Distributed Spectrum Clustering Method

Gang Chen, Dawei Zhao*

College of Humanities & Sciences of Northeast Normal University, Changchun, Jilin, 130117, China

*Corresponding author

Keywords: Hadoop-based Semantic, Big Data, Spectrum Clustering

Abstract: With the continuous development of the Sematic Web and the associated dataset project, semantic data in various fields is being expanded on a large scale. At the same time, there is a complex semantic correlation between these large-scale semantic data. The mining of these related information is of great significance to researchers. In order to solve the problems of traditional computing engine's computational performance and scalability in the large-scale semantic data reasoning, a Hadoop-based semantic big data distributed reasoning framework is proposed, and the corresponding property chain is designed. Prototype reasoning system to efficiently discover potentially valuable information in massive semantic data. The experiment mainly focuses on the semantic association discovery between the ontology in the medical and life sciences. The experimental results show that the inference system has achieved good performance--expandability and accuracy.

1. Introduction

With the continuous development of the Semantic Web, Web Ontology Language (OWL), which is built on the resource description framework (RDF), has been widely used in ontology modeling and reasoning in various fields, including life sciences, media information, semantic space-time data, social networks, etc., the semantic data of various fields also exploded. Taking the linked open data project as an example, it proposes the concept of linked data, which aims to call people to publish existing data as semantic link data, and Figure 1 links open data. The dataset map interconnects different data sources. As of September 2011, the LOD project contains a total of 295 data sources, 31 billion triples records and approximately 500 million RDF associations. Figure 1 shows the size of each data set in the LOD project and between contact. There are many hidden complex relationships between these massive semantic data. We can get the potential semantic information by inferring the existing semantic information. These hidden semantic relations are of great significance in practice. For example, biomedical workers can use semantic reasoning to derive drug associations to aid in the development of new drugs. Website data analysts can use user information to reason to discover potential relationships between users and users (such as the same preferences, the same interests).

2. Complex Association Semantic Big Data Reasoning

Traditional semantic data reasoning methods such as RDFS reasoning, OWL reasoning, and SWRL reasoning are all based on the established rules of their vocabulary. Take the rule 11 of RDFS semantic inference as an example. This rule: <x, rdfs:subClass, y> ∧ <y, rdfs:subClass,z>→<x,rdfs:subClass,z> indicates that if class x is class y Subclass, and class y is a subclass of class z, then we can get a new conclusion through this semantic rule that class x is also a subclass of class z. By analyzing these inference rules, we can see that they are generally only applicable to reasoning in a particular body. When we need to discover the complex association information between different domain ontology objects, this reasoning method is no longer applicable. . Complex relational semantic data reasoning refers to the use of semantic reasoning to

  DOI: 10.25236/dpaic.2018.034

mine the hidden related information for entities that may have associated relationships but are distributed among different domain ontology. The extraction of these related information is based on its semantic concept. The rules between the ontology in the model are indirectly inferred. For example, Chinese medicine experts now want to know the relationship between Chinese medicine and western medicine. We can express this process through a set of simple rules: Chinese medicine, treatment, disease, disease, possible drugs, western medicine, and Chinese medicine. Relationship, Western medicine. Through such a rule, we can relate the knowledge of two different domains of Western medicine and Chinese medicine. Of course, in the actual reasoning process, the association rules will show different complexity with different application of reasoning.

3. Distributed Reasoning Framework

Our distributed reasoning framework is mainly composed of three sub-modules: semantic modeling, rule extraction and parallel reasoner construction. The semantic modeling process is to construct a unified association knowledge by means of semantic and text processing from massive, different fields and different formats (obo, txt, xml, owl, dbms, etc.) from external heterogeneous environments. Map. The construction of this knowledge map requires large domain ontology to define the corresponding vocabulary, including classes, attributes, relationships, and so on. This domain ontology abstractly depicts the objective types or concepts and their attributes and relationships in the field. In the experiments in this paper, we designed a biotcmontology 1 across the Chinese and Western medicine field. Figure 7 briefly describes the basic objects and their semantic associations. With the support of this abstract model, we can determine the instance of the data that needs to be extracted. These different formats of data source processing methods are different, non-semantic data we directly convert it into a triple-format file through text processing, semantic data we convert its ontology file by calling JenaAPI and other semantic methods It is a triple format file, and all these triple files are imported into HDFS as data streams.

The rule extraction process is a process of extracting semantic inference rules from the unified link knowledge map according to the semantic concept model. The basis of semantic reasoning is the rule, because the core idea of system inference is to link the data of various fields through different associations, so that the Web can complete more intelligent association analysis or data mining through the navigation of these semantic rules. Rules are defined by experts referring to domain ontology and based on objective laws and facts. According to different application areas and inference objects, the rules we need to extract are not the same. In the following experiments, we consulted Chinese and Western medical experts and developed a chain of reasoning that captures the relationship between Chinese and Western medicine in accordance with the principles of biomedical internal laws. This chain will guide the reasoner in reasoning in the massive unified link knowledge map. Get the required associations.

The formulation of the inference join condition is critical because it determines the triplet pair that needs to be calculated for each iteration, thus controlling the total iteration flow and the number of iterations. The reason why the reasoning method in Fig. 5 is inefficient is mainly because its inference connection condition is harsh, and the connection candidate set corresponding to each iteration can only be in the first two types of entity triples (that is, the corresponding PID is 0 or 1) In the extraction, the other triples are not processed concurrently, so we need to design a more flexible connection strategy so that the system can infer as many results as possible in each iteration. Based on the above considerations, for the input instance triples, we have established the following connection conditions: 1) The PIDs of the two triples are adjacent. 2) The object of the triplet with smaller PID value is equal to the subject of another triple (each triple consists of subject, predicate and object). Suppose the system input is T0 <A0, P0, B0>, T1<B0, P1, C0> and T2<C0, P2, D0> in the example in Section 2.2. According to the connection rule, T1 is simultaneously Meet the connection conditions with T0, T2. So in the first iteration, we can calculate the connection candidate set as {<T0, T1>, <T1, T2>}, and infer the two triples to Tk<A0 in the first iteration. P0, P1, C0> and Tk+1 < B0, P1, P2, D0>. However, since Tk and Tk+1 no longer meet the connection conditions, the reasoning process ends. However, such reasoning results are obviously wrong. The

correct result should be Tk+2 <A0, P0, P1, P2, D0>. T1 cannot be connected to the two types of triples adjacent to the PID at the same time. To resolve this conflict, we introduced the third connection condition: the parity decision rule. For a triple with a PID value of k, if k is an odd number, then it can only be connected to a triple with a PID value of k-1. If k is an even number, then it can only be three with a PID value of k+1. Tuple join (here assuming k+1 is present). The above three connection conditions are used as the basis for the system to calculate the connection candidate set. Based on this connection method, the system can set all the instance ternary components into (n+1)?2 groups in one iteration task, then calculate the connection candidate set in each group instance, and then reason and predict according to the connection candidate set. The resulting result is again iterated as an input loop until the system infers the final result, so the time complexity of the entire algorithm is reduced to lbn.

4. Conclusion

Because the semantic reasoning tools running in the traditional stand-alone environment have bottlenecks in terms of computational performance and scalability, the proposed method of combining parallel technology and semantic reasoning has the ability to reason the massive semantic data, but they can only be processed in parallel. Inference or computational closure of simple associations within a single ontology cannot reason the cross-ontology data with complex semantic associations. This paper proposes a Hadoop-based semantic big data distributed reasoning framework, and designs a corresponding parallel chain inference prototype system based on attribute chain to solve this problem. The results of experimental reasoning time show that our inference system has good computational performance and scalability when dealing with large-scale complex association semantic data reasoning. At the same time, the correlation results obtained by verification and reasoning have also achieved good accuracy. This guarantees the effectiveness and availability of the system.

Acknowledgements

Fund Project:

1) This work was partially supported by The Education Department of Jilin province science and technology research project "13th Five-Year" Kyrgyzstan UNESCO Zi [2016] No. 159th (English translation)

2) Jilin Provincial Department of Education “Thirteenth Five-Year” Social Science Research Project, Ji Jiaoke Wenhe [2016] No. 506, Research on Software Professional Training Mode of Independent College

3) Jilin Province Education Science "Thirteenth Five-Year Plan" 2018 annual project, the project approval number is GH181034, the research and practice of the wisdom classroom teaching mode in the "Internet +" era.

4) Jilin Higher Education Society 2018 Higher Education Research Project, Question No.: JGJX2018D415, Research and Practice of Computer Teaching Reform under the Mode of "Internet +" Applied Talents Training

References

[1] Ge Leiwei.Design of intelligent large data storage architecture for power distribution [J]. Electric Power Automation Equipment, 2016, 36(6):194-202.

[2] Wang Zhiyong. The status quo and trend of precision medical development [J]. China Medical Devices Information, 2016, 22 (6): 5-10.

[3] Song Yi. Research and Application of Data Integration Technology for Integrated Distribution Network Planning and Design Platform [J]. Power System Technology, 2016, 40(7): 2199-2205.

[4] Zhang Xing. Analysis and Suggestions on the Status Quo of Power Network Disaster Prevention

and Reduction [J]. Power System Technology, 2016, 40(9): 2838-2844.

[5] Zu Xiangrong, Bai Yan, Yang Jiankun. Real-time monitoring and analysis of user demand response performance based on complex event processing [J]. Power System Technology, 2016, 40(10): 3220-3227.

[6] Long Suyan. Research and application of power settlement method based on metadata extension technology [J]. Power System Technology, 2016, 40(11): 3328-3333.