Distributed Storage of RDF Based on Clustering

Yonglin Leng, Fuyu Lu
College of Information Science and Technology, Bohai University, Jinzhou, 121000, China

***Abstract*—With the wide application of RDF(Resource Description Framework) data, the data volume grows rapidly. Therefore, RDF storage has become a hot research issue in data storage field currently. Distributed storage is an effective way to solve the storage and query of RDF data, and data partition is the premise of data distributed storage. In this paper we use graph clustering idea to realize the effective partition of RDF data. RDF can be described as a directed graph, so in this paper we use P-Rank (Penetrating Rank) algorithm to calculate the similarity of RDF graph node pairss, and then the improved K-means clustering algorithm is implemented to cluster the similarity results, so as to realize the distributed storage of RDF data. The experimental results show that, this method can complete the RDF data partition effectively, makes the intra-cluster similarity is smaller, and the larger the inter-cluster similarity.**

Keywords-RDF, directed graph, P-Rank, clustering.

I INTRODUCTION

RDF (Resource Description Framework) is a data model which was proposed by W3C for expressing metadata information about Web resources[1]. Two different forms was used to describe the RDF data, that is triple and directed graph. The triple takes the form <subject, predicate, object>, where the subject is the entity, the predicate is the attribute of entity and the object is the name of the entity or literal. In graph representation, each of subject and object is represented by a node, while a predicate is represented as a labeled directed edge from subject to object[2].

With the rapid development of the semantic web, the RDF data shows growth trend of massive. The storage and management by single machine has become the bottleneck of development of RDF data. Distributed relational and object data model was used to solve the scalability problem of RDF data, but they can not meet the low data redundancy and high query performance simultaneously. If the management of RDF data by graph model can not only avoid the conversion between RDF logical data and physical data model, but also can use the mature graph algorithm to optimize the RDF data query[3-5]. One of the key technology of RDF data distributed storage is the partition. Clustering is an effective method for graph partition. In order to clustering, we need measure similarity of node pairs. The similarity node pairs mainly be measured by the graph structure information [6] or graph node attribute and structural information[7].

P-Rank is an important measure model for structural similarity, which was widely used in data mining field, such as collaborative filtering, network graph clustering, KNN query. P-Rank(Penetrating Rank) algorithm was proposed by Zhao et al.[8] based on SimRank[9]. P-Rank overcome the deficiency of SimRank which only take the in-link relationship into structural similarity computation. In this paper, we use P-Rank algorithm to compute RDF graph node pairs similarity, then use improved K-means clustering algorithm to cluster the graph node, finally according to the clustering results to achieve RDF distributed storage.

II SIMILARITY MEASURE

SimRank is a simple and recursive algorithm which was used to compute the similarity between node pairs. The main assumption behind SimRank is that two objects are similar if they are referenced to similar entities[8].

Because SimRank only considered the in-link edge of node in the measurement of node pairs similarity, so the result existed a lot of 0 and also produced a large number of outliers in clustering.

P-Rank was proposed by Zhao et al. based on SimRank. P-Rank overcome the deficiency of SimRank which only take the in-link relationship into structural similarity computation. P-Rank contains two meanings:

i) Two objects are similar if they are referenced by similar entities;

ii) Two objects are similar if they reference similar entities.

Given a RDF directed graph $G=(V,E)$, where V is the set of vertices and E is the set of edges, For a node $v \in V$, we use $I(v)$, $O(v)$ as the sets of in-link and out-link of v separately, $|I(v)|, |O(v)|$ as the number of nodes in $I(v)$ and $O(v)$.

Let $s(u,v) \in [0,1]$ denote the similarity between two objects u and v, the iterative similarity computation equation of P-Rank is as follows:

$$s(u,v)=\begin{cases} \dfrac{\lambda \times C_{in}}{|I(u)||I(v)|}\sum_{i=1}^{|I(u)|}\sum_{j=1}^{|I(v)|}s(I_i(u),I_j(v)) + \dfrac{(1-\lambda)\times C_{out}}{|O(u)||O(v)|}\sum_{i=1}^{|O(u)|}\sum_{j=1}^{|O(v)|}s(O_i(u),O_j(v)) + & u \neq v \\ 1 & u = v \end{cases} \quad (1)$$

Where $\lambda \in [0,1]$ is the weight coefficient, C_{in} and C_{out} is the decay factor for in-link and out-link separately[9].

III IMPROVED K-MEANS CLUSTERING BASED ON SIMILARITY MATRIX

The basic idea of the K-means algorithm is to selected k points randomly, each point as a initial centroids and the other points is distributed to the nearest cluster according to the distance to each center. After that algorithm recalculate the average value of each cluster and update the centroids with the average value. This process is repeated until the criterion function E converge.

Instead of selecting initial centroids randomly, we follow the motivation of identifying good initial centroids from the degree point of view[10]. If the degree of a vertex v_i is large, it means v_i has more edges. In order to reduce the query communication of storage nodes, and therefore the node have more edges should be located in the clustering inside rather than the marginal.

We sort all vertices in the descending order of their degree values. Then select the largest k vertices from the sorted list as the initial centroids $C = \{c_1, c_2, \ldots, c_k\}$, assign other vertex $v_i \in V/C$ to its closest centroid according to the similarity between v_i and the centroids.

When all vertices are assigned to different cluster, the centroid will be updated with the most centrally located vertex in each cluster according to the formula (2). Where $\overline{v_i}$ is the average vector for cluster V_i [11].

$$R(\overline{v_i}) = \frac{1}{|C_i|} \sum_{v_k \in C_i} R(v_k, v_j), \forall v_j \in V \qquad (2)$$

Then the new centroid is computed by the formula(3)

$$c_i = \arg\min_{v_k \in C_i} \left\| R(v_k) - R(\overline{v_i}) \right\| \qquad (3)$$

The clustering process iterates until the clustering objective function E converges.

$$E = \sum_{i=1}^{k} \sum_{v_k \in C_i} \left\| R(v_k) - R(\overline{v_i}) \right\|^2 \qquad (4)$$

IV EXPERIMENTS

A Data set and experimental environment

In this paper, we chose the DBLP data set as the test data set, which includes 2555 articles and 6101 citation relations. In this data set it involves ten computer science fields, so there are 10 sub graphs and a RDF summary graph.

The experimental environment: Inter I3 processor, 4GB memory, Windows XP operating system, C++ programming language.

In order to verify the effect of P-Rank algorithm on the RDF data partition, we compared P-Rank with the SimRank algorithm, the algorithm of weight coefficient value is 0.5, the damping factor is 0.8. The experiment measure the partition effect from two aspects.

B Similarity number between node pairs.

First, we statistics the value of the node pairs in similarity matrix of P-Rank and SimRank algorithm, the node pairs are similar if $s(u,v) \neq 0$, otherwise they are not similar.

Figure 1 describes the number of node pairs which exist similar in 10 RDF sub graphs. As can be seen from the figure, because the P-Rank algorithm compute node pairs similarity from the in-link and out-link two way of transmitting information, but SimRank only consider the node in-degree in similarity measurement, so node pairs generated using P-Rank algorithm on the number of similarity was significantly higher than that of SimRank algorithm.

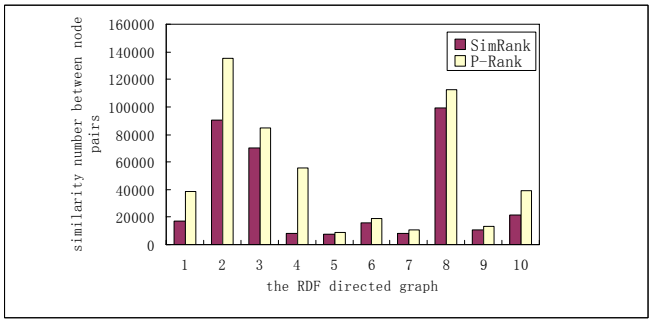

Figure 1. Comparison of similarity number between node pairs

C Cluster compression ratio.

We define the structure distance between two nodes as formula (5).

$$d(v_i, v_j) = 1 - s_f(v_i, v_j) \qquad (5)$$

Where $s_f(v_i, v_j)$ is the similarity of between v_i and v_j. Clustering compression ratio is described as follow:

$$C_f = \frac{\sum_{i=1}^{k} \sum_{x \in C_i} d(x, m_i)}{\sum_{1 \le i \le j \le k} d(m_i, m_j)} \qquad (6)$$

Where k is the number of clustering, C_i is the $i-th$ clustering, m_i, m_j represent the cluster center of C_i and C_j separately, numerator in formula (6) is the intra distance in clustering and denominator is the inter distance between clustering.

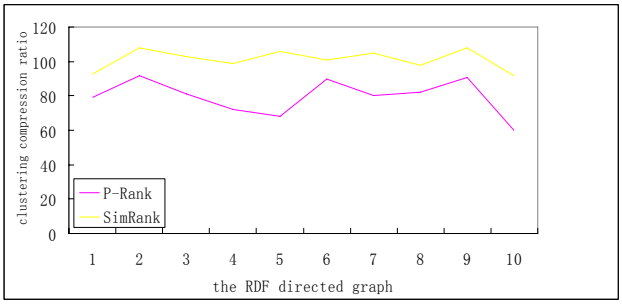

Figure 2. Cclustering compression ratio

Figure 2 describes the P-Rank and SimRank algorithm clustering compression ratio, as can be seen from figure P-Rank algorithm produced greater compression ratio, mainly due to the P-Rank algorithm allows more node pairs to exist similarity.

V CONCLUSIONS

The nature of the RDF data is a directed graph. In this paper, we realize the distributed storage of RDF data by clustering. A key problem of graph clustering is partition. SimRank and P-Rank algorithms are the structural similarity measure algorithm, so in this paper we chose P-Rank algorithm as the node similarity metric method, and then use the clustering algorithm based on the similarity matrix to achieve clustering segmentation of the RDF data, and then complete the RDF distributed storage of large data. The experimental results show that, this method can complete the RDF data partition effectively, makes the intra-cluster similarity is smaller, and the larger the inter-cluster similarity.

VI ACKNOWLEDGEMENTS

The research work was supported by Philosophy and Social Science of Liaoning under Grant No.L14AGL002.

VII REFERENCES

[1]RDF Primer.W3C Recommendation.Http://www.w3.org/TR/rdf-primer,2004.

[2]Wang Jin-Ling , Jin Bei-Hong , and Li Jing .AnEfficient Matching Algorithm for RDF Graph Patterns, *Journal of Computer Research and Development*. 2005,42(10):1763-1770.

[3]DU Fang, CHEN Yue-Guo, DU Xiao-Yong. Survey of RDF Query Processing Techniques, *Journal of Software*, 2013, 24(6):1222-1241.

[4]WU Gang.*Research on Key Technologies of RDF Graph Data Management*. Beijing: Tsinghua University,2008.

[5]Jiewen Huang,Daniel J.Abadi, Kun Ren.Scalable SPARQL Querying of Large RDF Graphs,*proceedings of the VLDB*.4(11),2011,1123-1133.

[6]H.Khosravi-Farsani, M.Nematbaksh, G.Lausen. SRank:Shortest paths as distance between nodes of a graph with application to RDF clustering,*Journal of Information Science*,39(2),198-210,2012.

[7]H.Khosravi-Farsani, M.Nematbaksh, G.Lausen. Structure/attribute computation of similarities between nodes of a RDF graph with application to linked data clustering[J]. *Intelligent Data Analysis*, 2013, 17(2): 179-194.

[8]P. Zhao, J. Han and Y. Sun.P-rank: A comprehensive structural similarity measure over information networks. *International Conference on Information and Knowledge Management*. 2009,553-562.

[9]G.Jeh and J.Widom.SimRank:a measure of structural-context similarity[J]. *In Proceedings of the eighth ACM SIGKDD conference(KDD'02)*.2002,538-543.

[10]A. Hinneburg and D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. *In Proc.1998 Int. Conf. Knowledge Discovery and Data Mining(KDD'98)*, 58–65.

[11]Zhou Y, Cheng H, Yu J X. Graph clustering based on structural/attribute similarities[J]. *Proceedings of the VLDB Endowment*, 2009, 2(1): 718-729.