Query Processing Based on Similar Nodes on SimRank Graph

LI Li
Dalian Polytechic University College of Arts & Information Engineering
116000

Abstract: With the exploding of information on Internet, relationship graphs have widely used in research areas, such as subgraph search, subgraph matching and subgraph mining. However, graph in real lives is always large-scale with millions of nodes and edges, leading to the analysis on the graph time-consuming and space-consuming. This paper analyze on the structure feature of graphs and discover the cluster in the graph. Based on the cluster feature of graphs, the author proposed a new algorithm which focuses on the similar vertexes query on the graph. At first, apply k-mean method to decompose the graph into small part, and then the similar degree on the part can be computed.

1. Introduction
With the proliferation of information on the Internet, the structure of graph data is widely used in real life and scientific research. Typical examples include personal relationship networks, partnership diagrams, protein structure information, molecular structure maps, and website link information. All representations of this information require a graph data structure. However, in real life, data information and daily drama have increased, resulting in a large scale of existing map data. In many applications and scientific research, we need to analyze the commonality between the graph data structures, and to analyze the characteristics of these commonality data, and then guide our future research.

At present, many applications are built on large-scale graph data structures, such as query recommendation, collaborative filtering, document similarity calculation, and fraud detection. These applications require node similarity queries on existing graph data structures. Similarity calculation is a hot issue. According to its application background, similarity calculation is divided into different aspects, such as document similarity analysis, link similarity calculation, and node recommendation in P2P environment. The analysis of these similarities can be basically divided into two aspects, namely, document similarity analysis and link similarity analysis in the network. The analysis based on document similarity can judge the similarity of documents according to the attributes of documents. The calculation of network link similarity mainly analyzes the similarity through links between web pages.

However, the analysis of these two aspects is difficult to meet different kinds of applications in various aspects, because there are many data structures in real life that are neither document type nor web page type, such as molecular structural features, so we need a universal and suitable calculation method to analyze the similarity between nodes. In view of this situation, we analyze the characteristics of existing graph data structures, analyze the similarity between nodes based on the link between nodes, and propose a universally applicable node similarity analysis algorithm.

The size of existing data is generally large, and in many cases, the analysis of node similarity in the graph can be very expensive. Considering the above points comprehensively, this paper proposes a
node similarity analysis algorithm based on SimRank in combination with k-means algorithm.

2. Related Work

At present, the similarity of nodes has been applied in many fields. The main research fields include subgraph query, subgraph matching, and subgraph mining [1]. Classical similarity analysis algorithms include PageRank, HITS, SimRank, and so on. The PageRank algorithm is a classic algorithm widely used in search engines. This is the famous Google application. PageRank uses a score-passing idea to pass the scores of the nodes on the graph to all nodes in the graph. The HITS algorithm analyzes the authority of a website from the link of a website into a web page and links it out, and also considers the authoritative differences of different web pages, making the obtained web pages more important.

However, there are still many shortcomings in these algorithms. The SimRank [2] algorithm proposes some improved ideas for these shortcomings. The researchers use the basic principles and ideas in graph theory and combine the basic logic principles propose the SimRank algorithm that which are easy to understand in real life. The main principle of the SimRank algorithm is based on the following basic idea: if any two nodes are connected together with similar nodes, then these nodes are similar. This simple idea is consistent with common sense and easy to implement. Based on this simple idea, SimRank is superior to many existing similarity analysis algorithms.

The SimRank algorithm only considers the overall structure information of the nodes in the graph and treats all nodes uniformly. It does not consider the difference of contribution values between different nodes. Later generation analysis; differences between different nodes and proposed new algorithms for these differences [5.6.7]. In SimRank++, the concept of "track" and the concept of "right" are proposed. The trace describes the commonality between two nodes. The trace is proportional to the number of neighbor nodes shared between the two nodes. The weight is the meaning of weight. In the graph structure, the weight represents the weight of the edge in the graph. After introducing these concepts that represent the differences between nodes, SimRank++ makes the node similarity calculations in the graph more accurate.

Combining these classical algorithms, people apply them in different research fields, and put forward many algorithms applied to subgraphs. By introducing the concept of subgraph isomorphism, the researchers proposed FSM through a certain greedy strategy in the decision tree, which is a new frequent subgraph mining algorithm. Other scholars have applied these problems to the uncertainty graph [3]. By introducing a new probabilistic model, the problem of subgraph query in uncertain graphs is solved efficiently.

The data structure of the graph is widely used. Many scholars apply the graph structure to the traffic network, and solve the sub-graph query algorithm using the relevant frequent closed graphs in the traffic network [4]. Firstly, the algorithm eliminates the redundant information in the graph structure by using certain processing, and then uses the tree structure to connect the frequent subgraphs. In this tree structure, the related subgraph queries are performed, which solves the subgraph query problem.

Although there are many researches on graph data and related algorithms, there are still many shortcomings in the existing algorithms. The main ones are as follows: (1) The scale of the graph data is large, and the similarity between nodes is performed on the graph. The calculation is expensive. (2) In the actual application, many graph data are dynamically changed, and related queries for these dynamic graphs have yet to be studied. (3) The efficiency of the algorithm associated with the graph has yet to be improved.

3. Problem Definition

The related algorithms and all the work in this paper are carried out on the graph data structure. The formal description of the graph data structure will be given in this section. The mathematical description of related problems related to similar node queries will also be described in detail in this section.
3.1 Entity Relationship Diagram Model

Entity relationship diagram is a definition in discrete mathematics. At present, entity relationship diagram has been applied in many research fields, and there are many formal definitions about the related issues of entity relationship diagram. An entity relationship diagram can be expressed as follows: Given a graph, \( G = (V, E) \), and in that \( V \) is a collection of entities in the graph, \( E \) is a collection of entity relationships in the graph. \( V \) can be used to represent real-life characters, nodes or web pages. \( E \) is a collection of relationships in an entity relationship diagram that is used to represent attributes of relationships between pairs of entities. Through the analysis of the pair \( E \), we can analyze the similarity between any two nodes and get a set of relationships between the nodes.

3.2 Similar Node Query Problem

Among the problems related to entity relationship diagrams, there are many problems that need to analyze the similarities between entities, find entities with higher similarity, analyze the commonality between entities, and lay the foundation for analyzing related problems. Similar entity query problems can be defined as follows: Given a graph, \( G = (V, E) \), any entity \( e \) in graph \( G \), the entities in the entity relationship graph are analyzed to obtain the entity \( e \) with the highest similarity to the entity and return the related entity.

4. Based on SimRank Similar Entity Query Algorithm

4.1 SimRank Algorithm

The SimRank algorithm was proposed in 2002 by Glen Jeh and Jennifer Widom of Stanford University [1]. This algorithm takes advantage of the idea of passing, that is, if two things are similar to some similar things, then the two things are similar. The core idea of SimRank algorithm is based on the recognized daily knowledge and has universal rationality. On this basis, SimRank can calculate the similarity between any two things in any structure.

Using \( a, b \) to represent the two entities on the graph, then the similarity algorithm between the nodes represented by the SimRank algorithm is calculated as follows:

\[
S(a, b) = \begin{cases} 
\frac{C}{|I(a)||I(b)|} \sum_{i=1}^{\|I(a)\|} \sum_{j=1}^{\|I(b)\|} s(I_i(a), I_j(b)) & a \neq b \\
1 & a = b
\end{cases}
\]

Wherein, \( S(a, b) \) represents the similarity between the representative entities \( a, b \), \( I(a) \) represents the inbound neighbor set of the entity \( a \), \( |I(a)| \) is the indegree of the entity \( a \), \( C \) is the damping coefficient, is a constant of 0 to 1, representing the attenuation factor in the process of similarity transfer along the directed edge.

4.2 K-means Clustering Algorithm

For large-scale graph data, analyzing the similarity between nodes directly on the dataset will result in high cost of time and space. Therefore, in this paper, the large-scale graph data is firstly segmented, and the partition is based on the partition. Node similarity analysis can reduce the time cost and space cost in data analysis, and thus improve the efficiency of the algorithm. At present, the algorithm research on the graph segmentation is relatively mature, representing graph-based graph data segmentation, graph data segmentation based on local optimization, and heuristic-based graph data segmentation. Researchers have proposed many classical algorithms for various research fields, and later generations have proposed many improved strategies based on these classical algorithms. In this paper, a classical graph data segmentation algorithm, k-means algorithm, is used. The k-means algorithm was first used in the field of data mining and became a classic algorithm in the field of data mining. The current k-means algorithm is widely used in other research, such as biomedicine, medical
research, image segmentation, and so on.

The main purpose of the K-means algorithm is to divide the data in the graph into different subsets, so that the similarity within the subset is the largest and the similarity between the subsets is the smallest. For this simple purpose, the k-means algorithm can solve the data segmentation problem in the graph well by several iterations.

The calculation process of the K-means algorithm can be described as follows:
(1) Select k nodes randomly as the initial cluster center.
(2) For each node in the dataset, calculate the distance between the node and the cluster center, select the cluster center with the shortest distance, and divide these nodes into the nearest center.
(3) Recalculate the cluster center for each clustering result and select the new calculation result as the clustering center.
(4) According to the obtained cluster center, recalculate the distance and divide the data set. If the result of the data set changes, repeat steps 2, 3, and 4. If there is no change, stop the iteration.

This problem involves the calculation of the distance of the nodes in the graph. In this paper, due to the particularity of the data set used, we can't project the nodes on the coordinate axes, so we can't calculate the European space between the nodes and the cluster center distance, where we define the distance is the shortest path.

4.3 Based on SimRank Similar Node Query Algorithm

In this paper, a new similar node query algorithm is proposed for large-scale graph data. Since there is a lot of redundant information in the large-scale graph data, it takes a lot of time and space to perform similar queries for all the data in the graph. Moreover, the results obtained by performing similar queries on all the nodes in the graph are mostly meaningless. In this paper, by studying the fast-structure nature of the data structure of graphs, it is found that the closely connected nodes in the graph are generally densely linked. In view of the real phenomenon, this paper uses k-means clustering algorithm to cluster the nodes in the graph, and then we perform node similarity query for the results of these clustering. The SimRank algorithm is used to analyze the similarity between these nodes. The implementation steps of the algorithm in this paper are described as follows:
(1) For the data in the graph, given the number k of clustering results, enter the data set to be processed.
(2) Using the k-means algorithm to perform segmentation processing on the nodes in the graph.
(3) For the query point to be processed, first query the collection it is in, perform similarity query processing on the collection where the query point is located, and return the result.

5. Experimental Result
In order to verify the validity of the proposed algorithm, this paper has verified it on different datasets. At present, there are many kinds of graph data. We select several representative datasets for related experimental processing. The details of the selected experimental data are as follows:

| Number of vertices | Number of sides |
|--------------------|-----------------|
| NetScience         | 1589            | 2742            |
| Cit-HepPh          | 34549           | 421578          |
| roadNet-PA         | 1088092         | 3083796         |

In order to verify the accuracy and efficiency of the proposed algorithm, this paper compares the time and accuracy of the original algorithm and the algorithm. For these problems, we conduct experiments and enumerate the results in the following chart.

|                 | Karate | Adjnoun | Cond-ma |
|-----------------|--------|---------|---------|
| SimRank         | 0.27s  | 0.553   | 0.927   |
| Our algorithm,  | 0.26s  | 0.529   | 0.865   |

From the data in the above table, it can be seen that the algorithm used in this experiment has a
reduced experimental time compared with the original algorithm. After analysis, the main reason is that the algorithm in this paper deals with less data.

6. Conclusion
On the basis of predecessors, this paper proposes a similar node query based on SimRank. In this paper, Li uses the K-means algorithm to segment the graph data. For the segmented data, we use SimRank to analyze the similarity between nodes. Thereby a similar node query algorithm is implemented. In order to verify the algorithm proposed in this paper, the experiment is carried out on several different data sets. The experimental results demonstrate the accuracy of the algorithm and its efficiency. However, there are also many problems in this paper. The accuracy of the experimental results of this algorithm needs to be improved. In the future work, we will focus on how to improve the accuracy of the algorithm.

References
[1] FSM - Frequent subgraph mining algorithm based on subgraph isomorphism and structural isomorphism. Zhou Yang, Wang Feng. Computer research and development. 2007, 44 (z3).
[2] Jeh G, Widom J. SimRank: A measure of structural context similarity[C]//Proceedings of International Conference on SIGKDD, Canada, 2002: 538–543.
[3] Uncertain efficient query processing in the graph database. Zhang Shuo, Gao Hong, Li Jianzhong, Zou Zhaonian. Journal of Computers. 2009, 32 (10)
[4] A graph based on frequent closed graphs contains a query algorithm. Li Xiantong, An Shi. Journal of Electronics. 2010, 38 (12)
[5] Antonellis I, Garcia-Molina H, Chang C C. SimRank++: Query rewrite through link analysis of the click graph [C]//Proceedings of International Conference on Very Large Data Bases, Auckland, New Zealand, 2008: 408–421.
[6] Michelle Girvan, M. E. J. Newman. Community structure in social and biological networks [J]. Proc.Natl Acad. 2002, 99:7 821~7 826
[7] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the Web[R]. Stanford Digital Libraries, 1998
[8] Lizorkin D, Velikhov P, Crrinev M et al. Accuracy Estimate and Optimization Techniques for SimRank Computation [C]In VLDB 2008: Proc of the 34th International Conference on Very Large Data Bases, 2008: 422-433