Journal of Education, Teaching and Social Studies
ISSN 2642-2336 (Print) ISSN 2642-2328 (Online)
Vol. 1, No. 2, 2019
www.scholink.org/ojs/index.php/jetss

Original Paper

Research on We—Media Information Retrieval Technology of Knowledge Map

Zhu Shanshan[1*]

[1] Department of Language Information Processing, Information Engineering University Luoyang Campus, 471003, China

Received: September 22, 2019 Accepted: October 18, 2019 Online Published: December 30, 2019

doi:10.22158/jetss.v1n2p130 URL: http://dx.doi.org/10.22158/jetss.v1n2p130

Abstract

As the development direction of information retrieval technology gradually evolves toward the relationship of search entities, traditional relational databases are difficult to satisfy, and graph databases are specifically created to handle the relationships between data. This article explains the basic concept of graph database, and takes the example of domain-specific database information retrieval as an example, analyzes its advantages and disadvantages, and analyzes the challenges faced by the graph database in full-text information retrieval.

Keywords

Graph database, relational database, information retrieval, large map data

1. Introduction

In the era of rapid development of information retrieval technology, each major information retrieval system in the Internet application industry can take the lead, not only because the retrieval system can bring a lot of data and knowledge, and people’s life and technology development is more and more inseparable from information retrieval technology. Most modern information retrieval systems are built on the basis of relational databases, and the search terms entered by users are used to return data resources such as web pages, pictures, audio and video that users need to query. However, due to the small amount of information provided by keywords, large query database and other factors, the information returned to users is still incomplete, inaccurate or slow query speed, which is also the future improvement direction of information retrieval technology.

On the one hand, the content stored in relational database is data rather than knowledge, which leads to the need for further analysis and processing of the results returned by the retrieval system. On the other hand, the larger the amount of data stored in the relational database, the more obvious the impact on the query speed. Therefore, considering the advantages of graph database in storing data, this paper analyzes the research of information retrieval technology based on knowledge graph.

2. Search System Development Trends and Data Gallery

2.1 Retrieval System Development

In the modern society with the explosive growth of knowledge volume, the development of retrieval system is no longer limited to the pages and pictures returned by simple search engines. People need to obtain more valuable knowledge more in line with individual needs from the retrieval system, instead of just the text content of a web page. Therefore, the development of retrieval system must pay more attention to knowledge mining, knowledge storage, knowledge representation and relational reasoning.

(1) Natural language understanding

User's understanding of input is an important aspect to improve the performance of retrieval system. Affected by the length of input and ambiguity of language, it is difficult to understand users' retrieval intention and the content they want to obtain.

(2) Knowledge discovery

Relational retrieval system still needs to solve the problem of relational mining. Nowadays, most of the network information exists in the form of text or audio and video, so it is difficult to dig out structured knowledge and the relationship between knowledge.

(3) Correlation analysis and reasoning

Based on the completion of the knowledge base, it is also very important to analyze and reason the correlation, which determines whether the retrieval results can meet the needs of users.

To sum up, users' demand for retrieval system will be more inclined to acquire knowledge, and relational analysis and reasoning can better meet such demand. Therefore, it can be said that the future development of retrieval system will be inseparable from the relationship.

2.2 Data Gallery

The world is made up of relationships, relationships exist in every corner of the world.

However, when dealing with relational problems, relational databases are not good at solving all kinds of relationships. A relational database is a database composed of a number of two-dimensional row and column tables that can be connected to each other. The first is that modeling is difficult, and uncomplicated data cannot be modeled to store data. Second, the performance is low. With the increase of the number of relationships and levels, more two-dimensional tables are needed, and the size of the database will increase, thus reducing the performance. Third, it is difficult to query. To query the relationship, multiple two-dimensional tables need to be joined together, while the JOIN operation will increase the query complexity. Fourth, it is difficult to expand. When adding new types of data and relationships, the database architecture needs to be redesigned, which will also increase the time to market of the database. It can be seen that relational database is not sustainable in relation processing.

In addition to relational databases, most of the data storage forms of non-relational databases are based on collections. Data are divided according to collections, such as documents in document databases, which makes it difficult to connect and establish relationships between data. To implement the foreign key functionality of a relational database, it is common for people to embed one data set directly into another data set to realize the dependency relationship between the two, but it is obvious that the creation of the relationship between the data sets requires significant storage overhead. The graph database was born to solve the relationship between the data.

A graph database is a non-relational database that USES the theory of graphs to store relation-al information between entities. It stores and queries data in a data structure called a graph, not a database of images. The most widely used graph database is Neo4j, which stores an entity as a node graphically at the bottom, as well as the relationships between nodes. By this way of storage, we can efficiently start from a node and find the connection between two nodes through the relationship between nodes. The two most important elements in Neo4j are nodes and relationships. When it comes to nodes and relationships, you need to introduce the Property Graph Model concept, as shown in Figure 1. In a graph-based database, a graph needs to record several nodes and relationships, which are used to associate two nodes. Both nodes and relationships can have their own properties, and nodes can be assigned to multiple categories.

In addition, the features of Neo4j&apos;s database give it many advantages. First, Neo4j sets up the relationships between nodes when it creates them, so it avoids problems that are difficult to deal with in complex query scenarios.

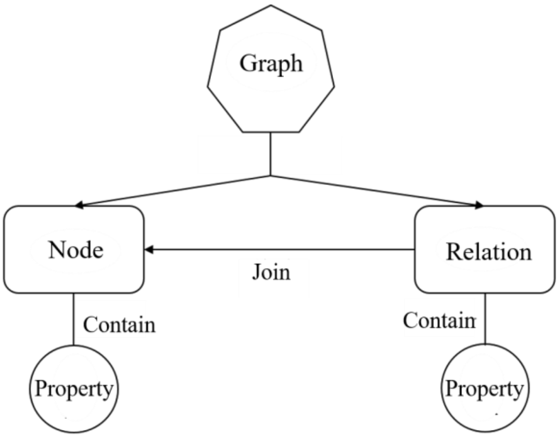

Figure 1. Attribute Diagram Model

Second, because the underlying database stores nodes and relationships directly as “graphs”, it is possible to keep the time complexity constant when querying a database node or relationship.

Third, Neo4j is based on the JVM implementation is easy to learn and debug, and you can get more code resources. Fourth, Neo4j provides an easy-to-understand query language, Cypher, and a built-in visual UI. Fifth, Neo4j has good support for ACID properties, as well as transaction mechanisms, so it retains the advantage of a relational database that handles transactions efficiently.

It can be seen from this that Neo4j graph database can not only store and query entities and relationships efficiently, but also retain many advantages of relational database, providing a good environment for data relationship processing.

2.3 Overview of Knowledge Map

The development of knowledge map has gone through three periods. The first period is the ontology period, which began with the emergence of the world's first expert system in 1968.

Expert system is a computer program that USES the knowledge representation and reasoning process stored in the computer to solve problems in the professional field (Lin, Wang, Zhou, Zhao, & Ma, 2017), which is the early form of artificial intelligence. Expert system is generally composed of knowledge base provided by domain experts and reasoning engine for analytical reasoning. By the 1980s, expert systems had proliferated. Representatives of this period are the Cyc project established by Douglas Lenat in 1984 (Yu, Zhang, & Gao, 2013) and the WordNet English dictionary project established by Princeton university in 1985 (Jiang, Peng, & Peng, 2015). The Cyc not only contains the ontology knowledge base that holds the massive knowledge, but also has the powerful deduction ability. WordNet, on the other hand, completes the compilation of an online dictionary, providing a powerful vocabulary for computers.

In 2001, with the emergence of Wikipedia, knowledge map entered the semantic web era (Judith, 2011). The emergence of Wikipedia promoted the construction of many structured knowledge bases based on Wikipedia. In 2006, Berners put forward the concept of Linked Data (Aleksey, 2016), calling on researchers to formulate uniform rules for knowledge Data release and make knowledge Data public on the Internet. The concept is proposed to combine the linked data to form the knowledge data of network structure. A representative project based on the linked data concept is DBpedia (Cihan & Adnan, 2018), which is also the first large-scale linked data project.

In the knowledge map project introduced above, the formation of knowledge basically depends on experts' manual compilation or extraction from structured data sources, while text data without knowledge expression structure is ignored. However, there is a large amount of unstructured text data in the Internet, which is also a high-quality information source. During the same period of the development of linked data, information extraction techniques for obtaining knowledge data in unstructured texts also developed. In 2007, Banko (Artem, Irina, Mansoor, Alexander, Christopher, & Charles, 2016) of the university of Washington first proposed open domain information extraction (OIE), which directly extracted the "head-entity-relation-word-tail entity" relational triplet from large-scale free texts (Jim & Sun, 2014). With the continuous progress of information extraction technology, Google Knowledge graph was launched in 2012 and entered the era of Knowledge graph.

3. The Design of "We Media" Information Retrieval Architecture Based on Knowledge Graph

As the data in "we media" is generally structured or semi-structured, and the data involved is small and easy to process. In addition, the retrieval system established in the professional field is easier to obtain the improvement of recall and accuracy, and the information in the database is more refined. Therefore, in this paper, we media information retrieval based on knowledge graph is taken as an example to elaborate its architecture design.

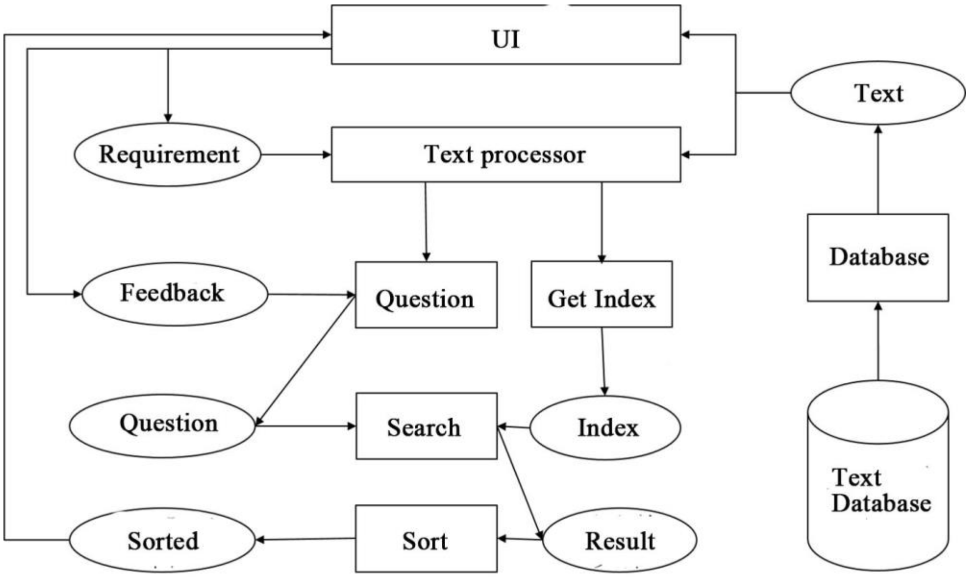

Figure 2. Information Retrieval Structure Diagram

The traditional retrieval system architecture based on relational database is shown in Figure 2. Generally, data is stored in a text database, and then the text data in the database is extracted from the words in the text, and then the text is indexed according to the extracted data. In the process of user retrieval, search the text according to the query words entered by the user according to the index database, then sort the retrieved text and return the sorted text to the user. In the following process, we also need to improve the retrieval results according to the user feedback.

In the “we media” information retrieval architecture based on knowledge graph (Figure 3), firstly, the data storage content is different from the general information retrieval, secondly, the data storage structure is based on graph theory, thirdly, the index construction is different, fourthly, the results and methods of retrieval feedback are different.

It can be seen that due to the storage characteristics of the graph database, this architecture requires high preprocessing of the data source. Since the graph database stores the attributes of the entities and the attributes of the relationships, the collected data sources need to be preprocessed before storage. Not all the collected data can be used directly, but it needs to be preprocessed structurally into stored nodes and relationship properties. Therefore, the content stored in graph database is very different from the general relational database, and the preprocessing of text is required.

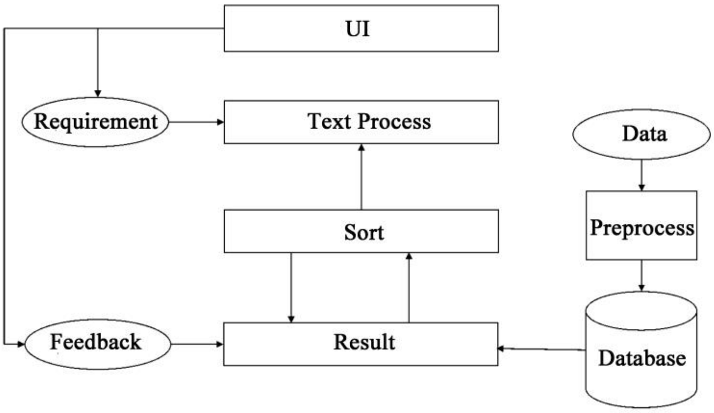

Figure 3. We Media Information Retrieval Architecture Based on Graph Database

The index construction of graph database is also different from that of relational database.

According to graph theory, graph data can be divided into several subgraphs, and then the contents of subgraphs can be extracted and summarized to form an index. However, in we-media information retrieval, data size is generally not large, so the index construction can be simplified. In this architecture, it is no longer necessary to sort the returned text, but to sort the nodes. The more important the most relevant nodes are, the more important the nodes are. Then, the ranking of the returned results is carried out through the relationship of the nodes and the weight of the relationship attributes as the associated nodes.

In addition to returning nodes and relationships, the returned results need to be texted so that they can be presented to the user in a way that is easy to read and understand. Because nodes and relational data are structured, they can be presented in a variety of ways, not just text but also diagrams and so on.

Next, according to the user feedback of the query results, the sorting results are optimized and the query results are added and deleted. Query result nodes with high user click rate should be given more weight and return more associated nodes, while query result nodes with low user click rate should be given less weight and less associated nodes. In addition, the schema needs to constantly update the contents of the graph database to meet user query requirements. Updating a graph database is more convenient than updating a relational database because the insertion of a graph is less expensive. Inserting a node or relationship into the graph database does not change the original modeling method, but only converts the data into the pre-established structure in the process of preprocessing, while updating the data in the relational database requires changing the database architecture.

4. Challenges

In the design of we media information retrieval, the use of graph database can bring many advantages. First of all, the results expected by users in the retrieval of the content of the subject area are knowledge content and related knowledge, while the graph database is generated by the analysis and search of the correlation. Second, the simple and efficient Chyphe is used as the query language, so the graph database can improve the query speed. Third, because domain knowledge data is basically structured or semi-structured, the attributes of nodes and relationships in the graph database are relatively complete and easy to read and understand.

However, for we media information retrieval based on graph data storage, although the mining of association relations and the query speed of entities will be improved, the following challenge-s still exist:

1. Integrity of data content. Since it is to retrieve the knowledge of “we media”, it is necessary to make requirements for the integrity of data content when the data volume is not large. The best case would be to include all the information about the domain, but this is difficult because of the intersection and complexity of the domain.

2. Real-time update of data. For the information retrieval of “we media”, the content of data should be updated in real time. Therefore, the operations of adding, deleting and modifying nodes and relations should be fully defined.

In addition to we media information retrieval using graph database, there are many researches on full-text data retrieval based on knowledge graph.

Because information retrieval technology is faced with huge data volume, the research on full-text information retrieval using graph database is still in the process of continuous exploration.

Large volumes of information must be stored in large graphs, which contain nodes and relationships in the billions. But at the same time, it can bring benefits to many applications, such as search engine development, e-commerce advertising push, path planning and so on. The management of large map data also faces many challenges, such as:

1. The data scale is huge and complex, including not only a large number of nodes and node attributes, but also complex association relationships.

2. Data flexibility is greatly increased, and the attributes of nodes and relationships are quite different due to the unstructured or semi-structured data contained. The heterogeneity of the data makes it difficult to store as a fixed schema.

3. Data is always under dynamic changes. As all kinds of data are constantly updated and changed, the contents stored in graph data are also changing all the time. The contents of changes include attribute changes of nodes and relations, addition and deletion of nodes, addition and deletion of relations, etc.

4. Large amount of data brings the complexity of query operation. Due to the large, complex and dynamic characteristics of graph data, it is difficult to query large graph data. The parallel distributed storage of large graph data makes the query first need the global graph information, search in blocks, and then query the subgraph data. Therefore, the establishment of large map index is also required.

Be worth what carry is, many scholars in order to solve these problem will figure with other relational database or combined use of relational database, in memory to use in the form of agraph database, query speed improved, at the same time also can reduce the complexity of the storage, so this method also has obtained the good effect.

5. Conclusion

This paper discusses information retrieval based on graph data storage, and analyzes the system design module by taking we media information retrieval based on graph data storage as an example. Then, the advantages and challenges brought by graph data storage in we-media information retrieval and full-text information retrieval are discussed. Generally speaking, the development of the world makes things more and more closely connected, forming many networks and bringing more connections. Therefore, in the future, information retrieval technology will need to provide more and more relational analysis for users, and the addition of graph database will better solve this problem. Although it is still difficult to store and manage graph data, its application in information retrieval is more important because of its increasingly mature technology.

References

Aleksey A. Mamchich. (2016). Models and Algorithms of Information Retrieval in a Multilingual Environment on the Basis of Thematic and Dynamic Text Corpora. *Cybernetics and Information Technologies*, *16*(1). https://doi.org/10.1515/cait-2016-0008

Artem Lysenko, Irina A. Roznovăţ, Mansoor Saqi, Alexander Mazein, Christopher J Rawlings, & Charles Auffray. (2016). Representing and querying disease networks using graph databases. *BioData Mining*, *9*(1). https://doi.org/10.1186/s13040-016-0102-8

Cai Xin, Ruan Yilong, & Shi Yirong. (2016). Design and implementation of IPTV content knowledge base based on knowledge map. *Telecom Science*, *32*(12), 32-36.

Cihan Küçükkeçeci, & Adnan Yazıcı. (2018). Big Data Model Simulation on a Graph Database for Surveillance in Wireless Multimedia Sensor Networks. *Big Data Research*, *11*. https://doi.org/10.1016/j.bdr.2017.09.003

Jiang Yang, Peng Zhiyong, & Peng Yuwei. (2015). Online genealogical record system based on knowledge map. *Computer application*, *35*(01), 125-130.

Jim Jing-Yan Wang, & Yijun Sun. (2014). From one graph to many: Ensemble transduction for content-based database retrieval. *Knowledge-Based Systems*, *65*. https://doi.org/10.1016/j.knosys.2014.04.003

Judith Winter. (2011). An Approach to XML Information Retrieval in Distributed Systems. *Methoden und innovative Anwendungen der Informatik und Informationstechnik*, *53*(4). https://doi.org/10.1524/itit.2011.0645

Lin Qisheng, Wang Lei, Zhou Xi, Zhao fan, & Ma Bo. (2017). Optimization and implementation of literature retrieval method based on knowledge map. *Microelectronics and computer*, *34*(10), 63-67.

Yu Suhua, Zhang Jun, & Gao Yan. (2013). Search algorithm research of embedded graph database. *Computer engineering and design*, *34*(12), 4204-4208.