PARTITIONING FRAMEWORK USING ONTOLOGY

Euiin Choi, WonSeok Kim and Byambasuren Temuujin
Dept. of Computer Engineering Hannam, University Daejeon, Republic of Korea

ABSTRACT

It's a major issue that how can retrieval worthy information in big data. Because big data can be used in company's success according how to take full advantage of big data analysis. Currently, search technologies about being stored distributed and duplicated data does not need to strong consistency. In order to provide more accurate results than the current keyword-based search, a logically expressed search technique using a knowledge base is needed. However, end users are unaware of the complex, formalized queries and schemas of the knowledge base. So the search system should be able to interpret the meaning of user keywords. So, in this paper, we are going to propose a new framework using an ontology.

KEYWORDS

Keyword, Knowledge base, Ontology, Big Data, Document Store

1. INTRODUCTION

Big Data is everywhere. According to the latest study by BT Germany, the data volume is expected to grow dramatically over the next two years. It will become reality in many companies have already started or are about to start to look into big data in depth. Search technology studies such as data envelopment analysis (Cooper 1984, Rijsbergen 1986a, 1986b) and R.Capurro information studies have been made. However, this search engine does not improve the quality of service. One of the reasons is its techniques still focus on only matching method index and query (keyword). So, distributed and duplicated data being stored are growing more and more interested in partition tolerance and weak consistency, availability. The retrieval is to collect / analyze / process the data, and then to provide the information that the user wants by using the stored data. The retrieval techniques can be largely divided into keyword-based and semantic-based retrieval technology. Recently, keyword-based retrieval technology has been developed and utilized in various fields. Semantic-based retrieval techniques have been developed for specific domains. However, as the development of metadata refinement and expression technology such as OWL (Web Ontology Language), which is an ontology expression language has been developed for several years. Therefore, this paper proposes a new framework using an ontology.

2. RELATED WORK

2.1 Format Data

2.1.1 Natural Language Retrieval

Natural language retrieval uses linguistic, semantic information of user's query which are written by natural language as much as possible into search throughout parsing method. It uses document clustering techniques, stemming algorithms, natural language processing technology for full-text mining with meta data.

2.1.2 Social Retrieval

Social retrieval is a one of the document retrieval techniques based on the text mining. The purpose is extracting useful information based on natural language processing techniques in text data. It can check relations between words in social media messages throughout network analysis, and also measure the intensity of interactions between those words. It might be used for exploratory data search range and result's purification and select on search interface.

2.1.3 Semantic Retrieval

Semantic retrieval interprets the search query and intent of the user to increase the accuracy of the results. QA is grasping user intent from user questions and providing the correct answer from the document to user. It utilizes web resources called RDF, term-dictionary and semantic web technology.

2.2 Unformatted Data

2.2.1 Graph Retrieval

Face book and Twitter utilizing mathematical graph theory analysis. It mainly used to find the user who acting like hub while measuring user's social influence based on social network connectivity structure and connectivity its strength. It is read consumer trends or patterns, and applies promotion. Face book's graph retrieval is straight through the process and combines data-driven techniques.

2.2.2 Big Data Search on NoSQL

Due to growing social networking services such as twitter and face book, this data scale shows terabytes and petabytes data. But existing search techniques have been handled only gigabytes scale information. So now we use CAP theory. It is agreeing with distributed potential and mitigates the data availability and consistency. NoSQL is data structure that is easy to search of large volumes of data. It has document, key-value, column data store structure and use inverted index method.

2.2.3 Column Store

Column stores also as a kind of key-value each record is independent. When the query pattern is subset of the whole record, it is possible to form column families in the same way. Cassandra can be applied. It made based on distributed architecture called Dynamo. It is to solve this problem that when the existing head node is down, the entire system also down. It can be stored more data no matter what data format.

2.2.4 Document Store

Document store is a special form of key-value store. You can look up the actual document by using a single key. You can look up some value in those fields. Thus, it easy to search when record is full and data is subdivided at the same time. It does not need to look all document or entire record. Document stores can immediately or in real-time analysis of dynamic changing data. In particular, Elastic Search and MongoDB have outstanding big data analysis ability in Hadoop eco system because of its outperforming.

2.3 Comparison

2.3.1 Elastic Search and MongoDB

Elastic Search is an open source distributed search engine. All of the fields in the document should be indexing by key. So it takes an advantage of part or all of documents retrieval by a single query. It provides scale horizontally and availability of the service by detecting and removing some node which has problem. Elastic Search runs on more than a half client node. If one of field including number of fields which have wrong analysis query or lowercase, the search performance may be degraded.

MongoDB is form of binary JSON, and it store data in document. It is corresponded a huge data store and retrieval by master-slave replication such as replication set and automatic failover. MongoDB is useful when user want define indexes or dynamic query of frequently changing data required. In addition, write lock is used when it manages update. Document-based partitioning is structurally simple, but if it has the more data, it returns the less exactitude.

2.3.2 Ontology Based Knowledge Based

Prerequisite of the semantic search is a well-established knowledge base. In the proposed system designed ontology schema which refer to the free base in order to construct the knowledge base. The selection criteria of reference model on ontology design was consider the knowledge representation range of schema, the knowledge representation method, and an abundance of knowledge. The free base can be expressed directly to using OWL (Web Ontology Language). And they designed to represent only knowledge of particular domain and used a different expression method.

3. PROPOSED FRAMEWORK

Figure 1. Framework

The ontology migration manager constructs a knowledge base by converting the multimedia content data constructed in the existing relational database into a triple according to a pre-designed ontology schema and transformation rules. The knowledge base manager identifies a knowledge object for the keyword through a term mapper and a knowledge base index of a knowledge base, the search keyword inputted from the user. Keyword semantic interpretation manager analyzes the inputted user search keyword and transmits it to the knowledge base manager, and recommends the semantic analysis result to the user. The domain search manager receives the result of the semantic analysis of the search statement, constructs it as a search statement suitable for each domain, and provides the result to the user by filtering, ranking, and dynamic classification. The search application manager receives a search keyword from a user, delivers the search keyword, and upon receipt of the search result, performs rendering and provides the user with a rendering result.

We should provide related search results associated with complex structure of content by combining metadata more transparently. Therefore, we suggest following architecture. Firstly, combine the advantage from existing method to term-based partitioning. Then it will have excellent adaptability and scalability of system.

4. CONCLUSION

We currently go toward to technical trend which is large stems called big data in a distributed environment such as variety of mobile applications, human and IoT. Thus, Information retrieval in between numerous data becomes a more important factor over its quality and reliability. So we analyzed MongoDB and Elastic Search and found its weak points. It was hard to find document included contents what you want to get using document-based partitioning. Thus, we proposed new framework which has only advantages of the exist techniques, and using term-based partitioning. So it can improve performance of retrieval. We expect that it will support intelligent retrieval.

REFERENCES

- Ed. B. Cronin, 2003. *Annual Review of Information Science and Technology*. Vol.37, Chapter8, pp. 343-411.
- Ji-Suk Kim, 2012. *Technical discussion and analysis of big data usage*. Korea University.
- Jun-Myeon Jo. et al, 2012. *Smart Media Search Technology Trends and Forecasts*. Korea Electronics and Telecommunications Research Institute.
- N.K Nagwani, S. Verma, 2011. *A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm*. In Intl.J. of Computer Applications pp. 75-87.
- Su-Jong Lim. et al, 2010. *Mobile Intelligent Search Technology Trends*. Electronics and Telecommunications Trends Issue.
- William Mcknight, 2013. *Information Management*. Morgan Kaufmann, pp. 97-109.