Research on Design and Implementation of Knowledge Map Storage Access System Based on Ontology Reasoning

Jian Li*, Yanzhou Ma

Luoyang Campus of the PLA Information Engineering University, Luoyang, China

*Corresponding author e-mail: maomaotftfn@163.com

Abstract. In order to support the intelligent combat capability of future wars, the network information system needs in-depth research and exploration in the field of artificial intelligence. This paper designs a distributed aggregation storage mode with the characteristics of balanced storage load distribution and local node aggregation storage under Big Table model. A distributed parallel query engine that uses the Group-By mode to distribute parallel computation query trees. A knowledge map completion method based on Bayesian inference is proposed. Bayesian probabilistic inference theory and RDF implication inference rules are used to jointly infer the potential relationships between entity nodes to predict the relationship between new nodes and original nodes, thus improving the mining efficiency of potential factors in the model and the accuracy of prediction of unknown relationships. Based on Big Table model, the entity sets stored row by row are evenly divided and stored through random prefix and pre-partition operations to realize load balancing. At the same time, random prefixes can be uniformly distributed to nodes for storing entities of the same type, and aggregated by entity category on a single node. With the continuous development of knowledge map technology, future knowledge map learning and reasoning technology can be integrated with new fields such as machine depth learning, cloud computing, block chain, big data, biological genetic engineering, etc. to play an important social value.

1. Introduction

Knowledge map is a directed graph structure, which describes the entities and concepts existing in the real world and their relationships [1]. Search engines can use knowledge maps to expand the semantics of query keywords, thus improving the search quality. As a typical application of education big data, adaptive learning has received extensive attention from educational science and technology circles at home and abroad [2]. The adaptive learning system makes full use of the technological advantages of machine learning and big data to build a truly personalized learning platform driven by digitalization. It is an important technical problem to design knowledge-based systems to provide solutions to those application problems that require expert knowledge. The acquisition, expression of expert knowledge and the composition and explanation of reasoning process are important technical problems. People's cognitive system uses knowledge to accomplish many different types of intelligent tasks, the core of which is an explicit and declarative knowledge system, which is supported by a conceptual system [4]. The concept system of human being is non-task specific and has strong universality, so there must be
some effective mechanism behind it to support such strong generalization ability [5]. An in-depth exploration of knowledge map data management can provide help in various fields such as search engine, machine learning, reasoning and artificial intelligence. The storage mode and query engine designed in this paper store the query knowledge map, as shown in the right figure, which can not only ensure the system storage load balance and node storage aggregation, but also make full use of the cluster performance to ensure the query efficiency of the system and make the storage query of the knowledge map have good horizontal scalability.

2. Design of Storage Mode for Knowledge Map Storage Access System

2.1. Selection of data storage system.
When testing the storage performance, MongoDB stores RDF triples in the form of key-value in the document. A row in the HBase table represents an RDF triplet, and each row has three columns to store the subject, predicate and object of the RDF triplet. This article will compare the storage time with 100,000 RDF triples, 1 million RDF triples and 10 million RDF triples. The comparison results are shown in Table 1 below. N/A in Table 1 indicates that the data storage time cannot be estimated.

|                   | One hundred thousand RDF data sets(ms) | Millions of RDF data sets(ms) | 10 million RDF data sets(ms) |
|-------------------|----------------------------------------|-------------------------------|------------------------------|
| MongoDB           | 5116                                   | 62868                         | 466920                       |
| Neo4j             | 5824350                                | N/A                           | N/A                          |
| HBase             | 5027                                   | 58335                         | 385370                       |

Facing the requirement of extensible knowledge map storage and query, this paper chooses HBase based on Big Table model as the data management system of knowledge map. In the face of large-scale knowledge, HBase based on LSM tree has faster storage speed, which is more beneficial to large-scale knowledge storage than other two databases. The construction of the knowledge map starts from the original data, extracts the knowledge facts from the original data by using some technical means, and then stores the knowledge elements into the data layer and the mode layer of the knowledge base [6]. In the knowledge map, each node represents the "entity" existing in the real world, and each edge represents the "relationship" between entities. Knowledge map provides the ability to analyze problems from the perspective of "relationship". In the face of large-scale data, the row key query of HBase can still reach millisecond level. Although faced with small-scale data, secondary and MongoDB have better query performance than HBase under single entity query and Neo4j query commonly used in knowledge map. Thus, knowledge system, as the core of human cognitive system to perform intelligent tasks, makes knowledge have higher recognition of intelligence than data and information. Systems with high knowledge content tend to solve problems more easily than systems with low knowledge content, and systems with reasonable knowledge storage structure tend to have lower redundancy [7]. The potential relationships between entities can be well characterized, and the relationships between entities can be expressed in the form of vectors. By comparing the similarity of vectors, the quality of the result graph can be deduced.

2.2. Design of distributed aggregation storage mode.
Facing the huge and increasing number of entities and relationships, the knowledge map storage access system needs to design an extensible storage mode to improve the storage performance of the system. The bottom-up construction method needs to extract entities (concepts) with high confidence from public data and add them to the knowledge base, and then manually construct the top-level ontology pattern. Considering that the current information, personnel, equipment and other business databases
have been designed according to the requirements of the business system and cannot be easily changed, the knowledge map of the network information system is based on the current business database. The knowledge map uses the method of "search+knowledge base" to organize massive network data in a reasonable and orderly way, to display the relationship between knowledge in the form of graphs, and to display knowledge to users in a more intuitive way [8]. It can describe abstract concepts of different levels and granularity and is a knowledge mapping to the objective world. However, this storage mode has the problems of repeated storage and wasting a large amount of storage space. At the same time, because HBase sorts the row keys in dictionary order, this storage mode is also easy to cause data hot spots and affect the storage and query performance. In addition, if there are tens of thousands of predicates, the table will reach tens of thousands, which is not conducive to management. These entities are merged into an entity object with a globally unique identifier and then added to the knowledge map.

![Figure 1 The relationship between knowledge map construction of network information system](image)

The distributed aggregation storage model designed in this paper only uses a large table to store all RDF triple sets, where each row stores all attribute values and relational objects owned by an entity, that is, a complete entity object. From the original data source, entities, attributes, and relationships among entities are extracted, and then ontological knowledge expression is formed. Entity extraction refers to the identification and extraction of meaningful quantitative phrases appearing in text, which is a key link in the information extraction stage. Using the method of machine learning, ontology and the relationship between ontologies are extracted from the existing structured, semi-structured and unstructured databases, and ontology meta-models are further formed by extracting metadata from the ontology knowledge base, as shown in Figure 1. The method extracts Horn logic representing semantic information of knowledge map through knowledge map rules, and injects potential semantic information obtained by materialization into a representation learning model; A series of discrete named entities obtained through entity extraction have little use value, and the relationship between entities needs to be extracted from the original data. When a large number of entity queries with the same initial letter are encountered, the queries may generate a large number of accesses to one or two nodes of the cluster, causing hot data issues and seriously affecting storage performance. A simple and effective method is to fully consider factors such as the reliability of data sources and the frequency of different information appearing in each data source, and to assist with manual proofreading to determine which category or attribute value is ultimately selected.
3. Design of Distributed Parallel Query Engine for Knowledge Map Storage Access System

3.1. MIQE, a query engine using distributed memory iteration technology.

The distributed memory iteration technology is adopted to design the query engine MIQE based on Spark calculation model. For complex queries, if only the filter method of HBase is used to query, it will query the corresponding query criteria in the way of full-table sequential scanning. In the face of massive data, the query performance of filter method is low, which seriously affects the user experience. The test item is taken as the research object, and the relationship between the potential characteristics of the subject and the reaction on the test item is taken as the research content. Based on this research, the project characteristic curve [9] is proposed. Ontology in ontology knowledge base is a clear formal specification of shared conceptual model and an abstraction of real entities, which provides a certain degree of logical description for the relevance of information in specific fields. Considering the hardware environment of the computer, the limitation of client requirements, and the convenience of software installation, upgrade and maintenance, the prototype system developed and designed in this section adopts the B/S architecture constructed by MVC mode. All Spark operations are memory-based, so Spark can avoid I/O operations on a large number of disks. 2. Spark is similar to mapreduce model, providing Map phase and Reduce phase, but Spark provides richer operators in the two phases, so theoretically Spark is much faster than MapReduce. The relationship between various responses and potential trait variables is provided. The trait level of the subjects is evaluated by observing the answers to the test questions, and then some predictions or decisions are made.

Bayesian formula is also called posterior probability formula and inverse probability formula, as shown in formula (1) [10]. Assuming that A is an event in sample space $\Omega$, $B_1, B_2, \ldots, B_n$ is a division of $\Omega$, if $A \subseteq \bigcup_{i=1}^{n} B_i$, $P(A) > 0$, and $i \neq j$, $B_i \cap B_j = \emptyset$, $P(B_j) > O(i = 1,2,\cdots,n)$, Then

$$P(B_j | A) = \frac{P(A | B_j)}{P(A)} = \frac{P(A | B_j)P(B_j)}{\sum_{i=1}^{n} P(A | B_i)P(B_i)} \quad (1)$$

Bayesian network is formally a Directed Acyclic Graph (DAG). Nodes represent random variables, directed edges between nodes represent conditional dependencies, and arrows point from parent nodes to child nodes. BN organically combines the expression and calculation ability of graph theory with probability theory, making it have flexible dependent topology structure in dealing with uncertainty problems, easy to understand and explain, obvious semantics and effective multi-information fusion.

![Figure 2 MIQE design scheme](image)

The designed query engine MIQE can seamlessly support the storage system proposed in the previous chapter that stores knowledge maps in a distributed aggregation storage mode. Specifically, the query engine design scheme is shown in Figure 2 above. The distributed parallel query system consists of a data processing layer and a data storage layer. Most of the massive educational resources
on the platform exist as independent individuals or are classified simply, which is easy to cause "network trek" and "cognitive overload" in the learning process [46]. So that information resources in specific fields can be organized into an organic whole on the framework of ontology description. The logic processing layer is used for data preprocessing and learning inference. The preprocessing includes using NLPIR to segment unstructured data in the book corpus, tagging part of speech, entity recognition and grammar analysis to obtain intermediate results. Facing the overall load balancing based on distributed aggregation storage mode design and the knowledge map storage system with local aggregation storage, the query engine needs to first load the data in the storage system into Spark memory by means of new APIH adopRDD and express it in the form of abstract set RDD, and then MIQE traverses the query tree generated by query conditions from bottom to top. The construction of ontology knowledge base is a dynamic process of continuous improvement and perfection. Some ontologies can be built by extracting from structured or semi-structured business databases. The retrieval ordering of ontology is to carry out preliminary word segmentation, data cleaning and word frequency statistics on the input data from the upper retrieval frame, and to analyze key words based on the ontology library, and to calculate the TF-IDF value (term frequency-inverse document frequency) of words in combination with the context analysis results of long tail search, and to sort the analysis results according to the importance of words.

3.2. IIQE is a query engine with inverted index technology.

A query engine IIQE is designed by using inverted index technology, which aims to use index scanning to query entities, reduce I/O of disks, and use coprocessor technology to speed up queries by using parallel scanning capability at the bottom of the HBase cluster. Based on the entities and relationships in the business database, the ontology knowledge and association relationships in the ontology knowledge base, and the real world abstract relationships and rules in the knowledge meta-model base, the entities (ontologies) and relationships are extracted according to the business needs to form the required knowledge map. Data units in ontology, such as concepts, definitions, and instances, can be stored as knowledge, and reasoning and expansion can be carried out through ontology rules. The update and maintenance of ontology correspond to the dynamic update of knowledge map and the completion of knowledge map. Then, a heuristic universal filtering rule is adopted to filter the candidate feature word set. Finally, the semantic dictionary is used to calculate the similarity between candidate feature words, and the candidate feature words are clustered to complete the automatic discovery of relationship types. At this time, each type is a relationship type for automatic discovery. Call the knowledge map generation service to form business entities, ontologies and their relationships related to business requirements organization to build knowledge maps, and carry out knowledge retrieval, reasoning and business planning based on the maps to push the results of user requirements. The query speed is accelerated by setting up inverted indexes on attribute values and relational objects of query entities and combining parallel query operations on the server side of HBase. In this scheme, the Observer framework of coprocessor is used to establish inverted index, and the Region is queried in parallel under the EndPoint framework using the statistical table operation of EndPoint framework for reference.

![Figure 3 Schematic diagram of calculating table rows in parallel by distribution](image-url)
Based on the above-mentioned improved distributed aggregation storage mode, the query engine IIQE uses the coprocessor's Observer framework to build inverted indexes. Different from using the client put operation of HBase to establish indexes, establishing indexes on the server side through the Observer framework can not only fully ensure the integrity and consistency of data, but also avoid the loss of indexes. Attribute value pairs are used to describe the intrinsic characteristics of knowledge entities. Knowledge entities form a network of knowledge structures through relationships. Knowledge context is an important factor to distinguish and identify different knowledge activities, generate knowledge requirements, and determine the types, scope and instances of knowledge associated with current business activities. It implements client code through protocol (protocol) definition interface and collects and merges data through RPC(Remote Procedure Call) communication. Before the new knowledge concept is introduced into the model, the data should be structured, and then it should be fused with the existing knowledge concepts of the knowledge map to remove overlapping or low confidence knowledge concepts and avoid data redundancy. Therefore, knowledge map can be regarded as an upgrade of ontology. The quality and scale of ontology data directly determine the performance of knowledge map. Figure 3 above. Users need to add their own defined counting operations to the server first. When performing the counting operation, the user calls the coprocessorService() method of HBase on the client, and then the RPC operation in the execCoprocessor() method can count the Region of each RegionServer of the table in parallel.

4. Summary
As intelligence is a necessary capability for future operations, the network information system needs to conduct in-depth research and exploration in the field of strong artificial intelligence in order to provide basic theory and support for various operational applications. This paper designs a distributed aggregation storage mode with balanced storage load, local node aggregation storage and a distributed parallel query engine that can query query trees in parallel to ensure the storage scalability and query scalability of large-scale knowledge maps. Based on the structural learning inference of Bayesian network, nodes and paths between nodes of Bayesian network correspond to entity nodes and relationships between entities of knowledge map respectively, thus using probabilistic inference rules for computational inference. In order to speed up the query, two distributed parallel query engines with different schemes are designed based on distributed aggregation storage mode: MIQE and IIQE. MIQE uses distributed memory iteration technology to filter and join entities represented by abstract sets of parallel queries. IIQE uses inverted index technology and combines cluster distributed parallel scanning capability to query entities in a merging manner. Both distributed parallel query engines meet the requirement of knowledge map query scalability. However, in the process of constructing and using knowledge maps, there are still a lot of technical problems in the automatic construction and updating of ontology. Based on larger-scale data, modeling and application are carried out, and the deficiencies of the original method are found and improved in the process, which is further applied from the aspects of scale and application effect.

5. References
[1] Zhang Dezheng, Xie Yonghong, Li Man, et al. Construction of ontology-based TCM knowledge map [J]. Information Engineering, 2017, 3 (1): 035-042.
[2] Liao Meihong. Design and implementation of Q & A algorithm in e-commerce field based on Chinese knowledge map [J]. Shopping Mall Modernization, 2019 (5): 34-35.
[3] Chen Qingyun, Cao Jianfei, Chen Ronghuan. Research and Practice on Construction from Thesaurus to Knowledge Graph [J]. Agricultural Library and Information Service, 2019 (1): 44-53.
[4] Zhang Zhongwei, Cao Lei, Chen Xiliang, et al. Review of Knowledge Reasoning Research Based on Neural Networks [J]. Computer Engineering and Applications, 2019 (12): 8-19.
[5] Pan Lihu, Zhang Jiayu, Zhang Shuai, et al. Construction of knowledge map in coal mine field [J]. Computer Applications and Software, 2019 (8): 47-54.
[6] Chen Yadong, Xian Guojian, Kou Yuantao, et al. Research on Construction of Knowledge Map of Apple Industry in China [J]. China Agricultural Resources and Regional Planning, 2017, 38 (11): 40-45.

[7] Hu Jie, Tao Hongcai. Design and Implementation of Domain Question Answering System Based on Deep Learning [J]. Journal of Chengdu University of Information Technology, 2019 (3).

[8] Wang Xiaoyun, Yang Ziyu. Construction of National Defense Science and Technology Knowledge Map Based on Scientific Research Ontology [J]. Chinese Journal of Medical Library and Information Science, 2018, 27 (7): 6-13.

[9] Wang Xin, Zou Lei, Wang Chaokun, et al. Review of Research on Knowledge Graph Data Management [J]. Journal of Software, 2019 (7): 2139-2174.

[10] Li Zhiming, Shen Zhou. Research Trends in the History of International Landscape Architecture Based on the Analysis of Scientific Knowledge Atlas——Taking Papers Published in the Journal of History of Landscape Architecture and Design Landscape (1998-2017) as the Object [J]. Chinese Landscape Architecture, 2019 (6).