Application Research of Machine Learning Method Based on Distributed Cluster in Information Retrieval

Xiaolian Li, Kunying Li, Dexin Qiao, Yu Ding, Daiming Wei
Computer Application Technology Department
PetroChina Research Institute of Petroleum Exploration & Development
Beijing 100083, China

***Abstract*—With the rapid development and comprehensive popularization of information technology, in order to reduce resource consumption and environmental pollution, all walks of life have made efforts to promote the construction of information technology in response to the call of the state to achieve paperlessness. When more and more information data is accumulated at the PB level per second, usually we use a kind of information retrieval method to filter the files in the database that quickly and accurately search the data file which users care. This paper proposes a retrieval cluster structure based on distributed network design and performs unsupervised machine learning on the unstructured data content according to the classification learning model, so that the child nodes have the function of automatically identifying the semantics of the article, then the second inverted index is obtained by feedback learning results that it was based on the original index . The practical application results show that the design model is feasible, not only significantly improves the retrieval effect and the accuracy of data screening, but also provides strong support for subsequent big data analysis.**

Keywords-component; Information Retrieval; Distributed Cluster; Machine Learning; category labels; Secondary inverted indexing.

I. Research Background

The rapid development of scientific research in various enterprises and industries has brought a wealth of knowledge and products in the professional field. More and more fields are involved in research, especially the wide application of interdisciplinary subjects is a complex knowledge application practice. The cooperation between a large number of professional disciplines and the sharing of international knowledge areas brings the storage and management of large unstructured professional data resources. The cooperation between a large number of professional disciplines and the sharing of international knowledge fields has created a huge problem of storage and management of unstructured professional data resources. The information retrieval method is to query and manage these unstructured data by using modern computer management tools, so that it can quickly and accurately locate the files that users care about, promote the digital transformation of the industry, and produce better efficiency and higher aim. In the current high-level combination of informatization and industrialization, the success of enterprises depends on their own construction of information technology, create a new development path that promotes industrialization with informationization and facilitate informationization with industrialization. Therefore, it is especially important and urgent to use information retrieval to manage the knowledge and data of the whole enterprise.

The focus of information retrieval is to evaluate the sensitivity and importance of the key words of the data, analyze the potential search intent and target range of the user, and accurately feedback the results. In the industry, the innovation and exploration of information retrieval has been continuously improved, and better results have been achieved. The analysis time has greatly improved the time and cost of user screening, but how to better identify the user's intention of searching and the decomposition of documents in the professional field has always been a problem left over at this stage.[1]An efficient and effective method of information retri eval based on multi-tuple rough set is created, this new method is a generalization of the ordinary Rough set model, which can obtain inexact query results through the multi-group approximate set. [2] Studying how to use the database query technology provided by China National Knowledge Infrastructure (CNKI) to analyze and calculate the data collected by CNKI according to the formula of journal evaluation indicators, and use the search to obtain the corresponding information. [3] To meet the increasing demand of accurate ciphertext retrieval in cloud environment, MRDI (Multi-Rationality for Dual-Indexing) ciphertext retrieval solution is proposed and designed. However, the above methods do not use the data in the field of professional knowledge as the source, and the data content is effectively split and parsed according to actual request, so that the computer can effectively understand the key intention of the user retrieval. Therefore, this paper proposes a second inversion indexing technology based on category tags, which has more efficient and accurate information retrieval capabilities, integrates existing word segmentation techniques, and is optimized with the mature periodic data heat and category tags. The algorithm solves the problem of semantic understanding, retrieval efficiency, storage of massive unstructured data management capabilities and deep mining of all data in the information retrieval. It also provides optimization and management capabilities for category tags. The ability to continuously upgrade learning prototypes in the

978-1-7281-3681-3/19/$31.00 ©2019 IEEE
DOI 10.1109/CISCE.2019.00097

later stages is realized. The practical application results show that the construction scheme of the information retrieval is feasible, which significantly improves the machine's ability to understand and retrieve unstructured data.

II. Retrieval Cluster Structure Based On Distributed Network Design

As the size of the user query and the number of documents collected and collected continue to increase, a single server can no longer support the operation of the service, resulting in frequent service failures. [4] The solution currently provided is to introduce a copy mechanism and establish a multi-point cluster environment for data storage and analysis. Storage mode introduces a distributed concept that is widely used for the storage and access of big data. This technology greatly improves the overall performance and efficiency of the network, reducing the cost and consumption of network storage.

According to the [5] general distributed environment construction method, the two-node mode of the primary zone and the replica zone will be set; the primary zone node will monitor all the partition replica nodes, and if a certain partition node continues to be inaccessible within a certain period of time, then this node is marked as a unknown error and the host will rebalance the cluster. The main job of the replica partition is to be responsible for fault tolerance and undertake additional load requests. In order to avoid redundant data, the partitions on other nodes will automatically assume the work of the lost nodes, and temporarily remove the wrong nodes from the cluster to ensure the normal operation of the storage application environment. Therefore, each document library will be built an index structure, and create multiple shared work slice areas for each index. Each work slice area is a minimum work unit which carries a part of data. At the same time, the host will automatically load balance the workspace when adding and deleting nodes. This way significantly improves data security, resource availability and service efficiency.

In this research, in order to enable each child node to have the ability to analyze and understand data independently of neurons, realize the automatic classification, aggregation and feature recognition of data content by nodes. A search analysis instance is deployed on each data collection unit to provide complete content semantics for automatic analysis and processing of events. At the business-oriented service level, the state of using a single master node to process all files in a traditional search is changed, and the way of simulating a neural network gives the retrieval cluster a processing mode for the semantics of the data content.

III. Using Classification Learning Prototype To Realize Unsupervised Machine Learning

In the actual research process, it is found that the professional terminology and academic content for various industries are not satisfied by the popular semantic understanding algorithms at this stage. Taking the energy industry as an example, most users are more concerned about how to quickly retrieve the industry-related content what they need. On the contrary, it is not so sensitive to how to understand human emotions and thinking. In order to solve the above problems, this paper proposes to build a category label model based on the [6] distributed environment construction, which is applied to the retrieval analysis examples of each child node, enable each child node to automatically analyze and process the collected professional data file content, generate an index from the unsupervised machine learning from the title, abstract, body content and label key index respectively and feed back to the host. The machine learning architecture based on distributed environment is divided into three layers: regional slave layer, regional host layer, and Master host layer. The specific design of each layer is as follows:

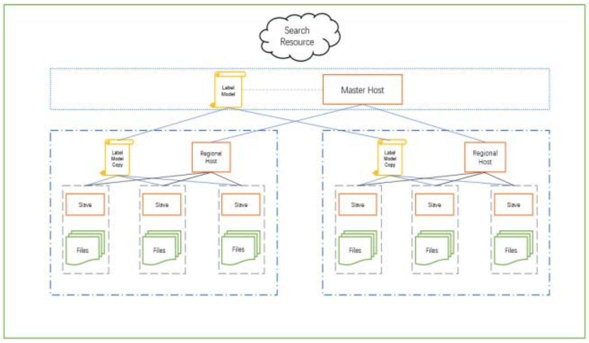


Figure 1. Machine learning architecture design based on distributed environment

1) Regional slave layer: In technical perspective, this layer is divided into two layers: data collection and data analysis. Data collection uses FileBeat to collect and classify files in real time, including different types of files such as doc, pdf, and excel. Then, different files are transferred to Logstash one by one for queue waiting. Data analysis is to parse the collected files. According to the set label learning model, the contents of the file are cut and split to match the corresponding labels, and the real intelligent semantic understanding is performed, and the feedback is summarized to the upper host.

2) Regional host layer: It mainly refers to the management and data collection of the slave nodes in a certain area and summarizes information to the upper layer. In addition, it is responsible for load balancing of the replica node, environment allocation service, and cleaning and classification of the collected data.

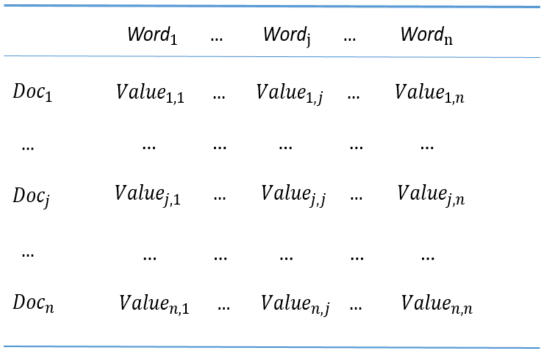
3) Master host layer: Unified management of each regional host, collecting all data for secondary combing, combining the weights of each tag in the tag model, and using the retrieval algorithm based on periodic data heat and category tags to score the data index. This approach makes the system closer to the user-oriented evaluation criteria for data scoring, and further optimizes the search accuracy.

This distributed machine learning framework design firstly automatically learns the prototype samples to locate the category keywords of the data, extracts the category keywords in the data and assigns corresponding labels, such as: oil production, drilling, fracturing, oil and gas, etc. Forming a document category label set, combining the obtained user recent query task target result list, calculating a correlation between the query result list and the label set. Secondly, the number of visits and the number of hits is combined to calculate the hit frequency, and finally the order of the feedback results is adjusted according to the above additional score. This architecture design can greatly reduce the I/O pressure brought by big data traffic. It has already given the learning ability of single-node preprocessing, and adjusts the learning samples according to the environment, reducing the large amount of consistency cost brought by artificially identifying content semantics.

IV. Key Technology Implementation

A. *Inverted Index Technology*

The main ability of indexing is to be able to locate quickly when retrieving a large number of data groups [7]. When large amounts of data need to be retrieved, if traditional linear matching is used to write data into computer memory, then match term in resource, so the consumption of time and money is very huge. In order to solve this problem, using a method of indexing documents or words, through the retrieval of the index list to achieve rapid positioning. It is necessary to generate a corresponding mapping between the association between text and documents in the process of building a search engine based on it, which can be divided into a positive index and an inverted index.



	$Word_1$	…	$Word_j$	…	$Word_n$
Doc_1	$Value_{1,1}$	…	$Value_{1,j}$	…	$Value_{1,n}$
…	…	…	…	…	…
Doc_j	$Value_{j,1}$	…	$Value_{j,j}$	…	$Value_{j,n}$
…	…	…	…	…	…
Doc_n	$Value_{n,1}$	…	$Value_{n,j}$	…	$Value_{n,n}$

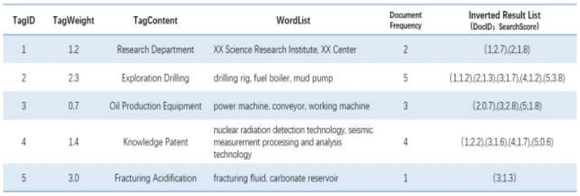
Figure 2. Positive Index

However, in the daily data retrieval process, the number of documents in the database will far exceed the number of characters recorded, so our search engine design will use [8] matrix inverted technology, then, an inverted matrix with keywords as the core is formed. Each keyword Word corresponds to the associated document Doc, so that the number of entries of the index is also much lower than the positive matrix. By inverting, the words we retrieve can quickly locate the associated document list in the index, [9] the important calculation factor of content relevance (Search Score in the figure below) is recorded in the inverted list, which facilitates the subsequent ranking of query results for effective scoring calculation.

B. *Secondary Inverted Indexing Technology Based On Category Labels*

In the course of this research, the machine's understanding of the content of the file is presented in various forms by various tags and word segmentation structures. Not only it carries a large amount of file data, but also the content of unstructured associated attachments is diverse. From the comprehensive analysis of the future and the efficiency and value of data mining, this paper proposes a secondary inverted indexing technique based on category labels. More efficient and accurate information retrieval capabilities, integration of existing word segmentation techniques, and sophisticated search optimization algorithms based on periodic data heat and category labels. It solves the problem of semantic understanding, retrieval efficiency, storage of massive unstructured data management capabilities and deep mining of all data in the future.

Taking the category label model as the core, the inverted index mentioned above is processed twice, in which a label layer is encapsulated based on the vocabulary, the label and the keyword are effectively associated, and then the second inversion processing is performed. Generate a second inverted index based on category labels. Then this technique is used to retrieve, drill down, and effectively score various types of unstructured data in the database. The results of the second inverted technique used by the search engine are shown in Figure 3.



TagID	TagWeight	TagContent	WordList	Document Frequency	Inverted Result List (DocID： SearchScore)
1	1.2	Research Department	XX Science Research Institute, XX Center	2	(1;2.7),(2;1.8)
2	2.3	Exploration Drilling	drilling rig, fuel boiler, mud pump	5	(1;1.2),(2;1.3),(3;1.7),(4;1.2),(5;3.8)
3	0.7	Oil Production Equipment	power machine, conveyor, working machine	3	(2;0.7),(3;2.8),(5;1.8)
4	1.4	Knowledge Patent	nuclear radiation detection technology, seismic measurement processing and analysis technology	4	(1;2.2),(3;1.6),(4;1.7),(5;0.6)
5	3.0	Fracturing Acidification	fracturing fluid, carbonate reservoir	1	(3;1.3)

Figure 3. Example of a second inverted index based on category labels

V. Summary

In order to achieve faster and more accurate search and feedback to the user's most concerned data files, this paper applies the architecture of the search engine based on distributed cluster design, and machine learning of the initial prototype using unsupervised learning for scattered child nodes on this model, and then the data association resource of the search engine is established by using the second inverted index technology based on the category label, and then the basic information of the unstructured data, the label information and the output of the final scoring algorithm are effectively correlated. So, it generates an index collection of data file groups that enables efficient classification and efficient querying of unstructured data in large-scale data sources.

References

[1] Zhifeng Ma, Hanchen Xing, Xiaomei Zheng. A Multi-Tuple Rough Set Approach for Information Retrieval[J]. Journal of Southeast University (English Edition),1999(01):63-68.

[2] Wenliang Xie, Yanmin Li, Yijun Zhang. Applying information retrieval technology in analyzing the journals[J], 2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies,2013(09).

[3] Junliang Tu, Bo Lei. Ciphertext Retrieval Technology of Keywords Semantic Extension based on MRDI [J]. Communications Technology,2017,50(12):2828-2832.

[4] Weiru Feng. Massive Unstructured Knowledge Management System based on Text Clustering Technology under Distributed Computing Environment [D]. Nanjing Aerospace University,2012.

[5] Aaron Kimball, Sierra Michels-Slettvet, Christophe Bisciglia. Cluster computing for web-scale data processing[J]. ACM SIGCSE Bulletin. 2008 (1)

[6] Sogrine M,Patel A. Evaluating Database Selection Algorithms for Distributed Search. Proceedings of the ACM Symposium on Applied Computing. 2003

[7] Zhang Y, Li J. Research and Improvement of Search Engine Based on Lucene. Intelligent Human-Machine Systems and Cybernetics,2009. IHMSC09. International Conference. 2009

[8] Libo Jia, Xiaoming Jiang, Qing Ye, Zhanfang Chen. A Frequent Item Set Mining Method Based on Inverted Index [J]. Journal of Changchun University of Science and Technology (Natural Science Edition),2019,42(02):117-119+124.

[9] Kunying Li, Dexin Qiao, Xiaolian Li, Yu Ding. Analysis and optimization of screening algorithms for unstructured data[J]. Advances in Computer Science Research,2019(04).