Application of Big Data Computing Model Based on Semantic Data Analysis Technology

Ganglong Fan1,* and Jian Shen2

1Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce, Luoyang Normal University, Luoyang, China
2Higher Education Admissions Office of HENAN Province, Zhengzhou, China

*Corresponding author e-mail: 85660190@qq.com

Abstract. Big data computing model refers to various high-level abstract models extracted and established from the diverse big data computing problems and requirements according to big data's different data characteristics and computational characteristics. In order to improve the data quality, a method of adding semantics to the data is proposed to help users identify big data's pattern. In this way, we can make use of the semantic value of the data, detect big data, and propose a big data computing model through the semantic analysis of the data. The paper presents research on big data computing model based on semantic data analysis technology.

Keywords: Big data computing model, Semantic data analysis, Big data, Semantic similarity, Text processing

1. Introduction

In the process of joint or join operation, it is necessary to distinguish synonyms from synonyms in order to complete semantic data merging. None of the tools that are now commonly used take semantic aspects into account, but only grammar. The difficulty now is for semantics to improve the quality of data. Misunderstanding of the data patterns will be a major obstacle to our success in solving the problem when developing strategies to correct the errors in the data.

Semantic similarity is the degree to which two words can replace each other in different contexts without changing the syntactic and semantic structure of the text [1]. To migrate the original system and application to the new distributed architecture, we need to make a great adjustment from the underlying platform to the upper application. Especially in terms of database schemas and application programming interfaces. Therefore, data quality is particularly important in big data environment.

Text processing is a word that tells computers how to understand human beings. High quality semantic analysis algorithms are usually modeled on a good set of text processing mechanisms, the two most important of which are Chinese word segmentation and word segmentation weight (Term Weighting). When you get a piece of text, how does the computer know its grammar and word composition? Referring to the response of the human brain to language, in general, word segmentation should be made first. Word segmentation is the reasonable separation of a paragraph of text into several root words to match the reference thesaurus.
Causality is an important part of mathematical statistics, based on perfect mathematical theory, it represents a regression model, and correlation is an important part of data mining, based on strong machine computing ability, represents neural network, decision tree algorithm, which makes people do not need to understand the complex causality logic can also obtain good analysis and prediction results.

For semantic data processing, we propose to give each data source, an error report, updated logs, and new semantic structures that use metadata [2]. Because of the application of cloud computing mode, distributed technology and cloud database technology, we do not need such a complex model; we can handle big data without considering complex computing algorithms.

High speed describes the speed at which data is created and moved. In the era of high-speed network, it has become a popular trend to create real-time data flow through high-speed computer processors and servers based on software performance optimization. Enterprises need to know not only how to create data quickly, but also how to process, analyze and return it to users to meet their real-time needs.

2. Big data's Computing Model and system Analysis

The metadata in big data is less than in traditional databases, and when data scientists want to analyze these data, they will encounter trouble, and the use of data quality management tools will help data scientists identify data types.

For each column in the data source, there are at least two categories, and to determine the main class, select the larger category. The ratio is calculated on the basis of the number of correct values belonging to this category. If the two classes have the same proportion, select another sample from the data source and analyze it with semantic data.

For the big data platform, you have mastered a lot of knowledge and skills, build Hadoop cluster, collect data on Hadoop, and use Hive and MapReduce to analyze data, synchronize the analysis results to other data sources [3]. Although there are tools that enable s1 and s2 to merge, this is semantically meaningless.

With the development of technology, mobile devices such as smartphones and tablets integrate more and more sensors, and their computing and sensing capabilities are becoming more and more powerful. Under the background that mobile devices are widely used, group intelligence perception has become an application hot spot in the field of mobile computing. A large number of users use mobile intelligent devices as the basic nodes to cooperate through Bluetooth, wireless network and mobile Internet to distribute perceptual tasks, collect and use perceptual data, and finally complete large-scale and complex social perceptual tasks.

Syntax invalid values are easy to find, can be judged using regular expressions, and processed results can be used to enrich the data dictionary, because there may be multiple languages in a column, as is shown by equation(1) [4].

\[ a_n(l+1) = a_n'(l+1) + \sum_{g=1}^{n-1} a_g(l) + a_n'(l+1) \]  

(1)

In formula 1, you can see that there are several column items in the data source l, and a is recorded as (Coli, k = 1-7). In V, observing the fourth column, "Beijing" and "London" are valid in syntax and semantics, but "pekin" and "Londres" are valid in grammar and invalid in semantics.

There are three methods to calculate the semantic similarity of words: the method based on knowledge system, the method based on corpus and the method based on network. Most of the methods based on knowledge system are based on WordNet. Wordnet is a semantic dictionary, which groups words according to the meaning of the entry, each group of words with the same meaning is called a synset (synonym set). Wordnet provides a short, summary definition for each synset and records the semantic relationship between different synset.
However, although the N-gram language model is simple and effective, it also has limitations [5]. Because only the position relationship between words is considered, there are still obvious deficiencies in interwork similarity, grammar and semantics, and the more complex neural network language model just fills this blank. The neural network language model uses each word $m (n \leq 1)$, $m (n \leq 2)$ in N-gram. $M (n \leq 1)$ is mapped to the word vector, and then the vector combination of each word is combined to form a larger vector as the neural network input, and the output is that the similarity between $P (mn)$, words is finally represented by the word vector.

The data mining theory has obtained the geometric multiple growth of the amount of data and processing ability, and many ideas and methods that cannot be verified can be realized. For example, the traditional BI analysis has a "centralized" step, that is, it is necessary to extract and centralize a large number of data before the analysis to form a complete data warehouse, which often becomes the bottleneck of the whole process of BI analysis [6]. However, the BI analysis based on big data distributed technology does not need to be "centralized", which greatly improves the level of agility and intelligence, thus promoting a major breakthrough in machine learning, semantic processing and other fields, and directly contributed to the advent of a number of commercial products, such as Mahout machine learning algorithm set, Siri voice assistant and so on.

Before big data, we usually divided data storage into relational database and file server. For unstructured data that will account for more than 80% of the total data, although NoSQL data storage has the advantages of scalability and availability, it is beneficial to trend analysis and provides a preliminary solution for big data storage.

As a result of the combination of multiple types of data, S can also be seen as a collection of strings, where the content is separated by semicolons and represented by column entries, a data schema for each record S.

In the process of enterprise information construction in our country, the phenomenon of block segmentation and Information Island is common. The same industry, such as transportation, social security system and so on, is also divided according to the administrative field, so it is very difficult to exchange and coordinate information across regions [7]. Seriously, even in the same unit.

Hadoop can meet hundreds of MB per second of log data acquisition and transmission requirements, and upload the data to the Hadoop central system. Public transport management through big data will face how to open public transport data, how to protect personal privacy, how to access traffic data, and so on, which can be protected by attention to personal private information. Most feedback time requirements are less stringent.

3. Semantic data Analysis process based on ontology and knowledge Network

Error reports include multiple exceptions in the data source: more than one category and language appear in the same column, inconsistent data formats, copies, and null values. Word vector space model is one of the most widely used strategies for calculating the similarity of words based on statistics. Big data will continue to become an important asset for all kinds of institutions, especially enterprises, and become a powerful weapon to enhance the competitiveness of institutions and companies.

Computing processing layer, such as hadoop, MapReduce and Spark, and various computing paradigm on top of it, such as batch processing, flow processing and graph calculation, including computing models derived from programming models, such as BSP, GAS, etc. The updated log is a collection of update behavior that is used for data sources such as: translated language, homogenized format.

Data analysis is the first step in the data processing process. It is also a quantitative analysis, including narrative analysis, such as the definition of schema, table, and domain and data source [8]. Data quality management tools, the current data analysis tools provide statistical data analysis, and do not solve the semantic analysis of data. This is used to extend the semantic indicator of the analysis process, as is shown by equation (2).
\[ p_{ij} = P(M_j(k) \mid M_i(k-1)) \] (2)

In formula 2, \( p \) is a variable parameter, which means the distance value of words. Similarity is defined as a real number between 0 and 1. When the two words are exactly the same, the similarity is 1; when they are completely different concepts, their similarity is close to 0. The calculation methods of semantic similarity are based on knowledge system, based on large-scale corpus and based on network.

For a text information, some text processing methods in semantic analysis begin to perform higher-level semantic analysis tasks after word segmentation and Term Weighting scoring. At present, the main body model used in the industry is LDA, whose algorithm execution efficiency is high, which can solve the problem of keyword topic association.

This is to calculate the semantic similarity by taking the edge as the distance. If all the edges in the tree semantic network, that is, the branches of the tree, are equal in length, then the number of edges can be used as a measure of distance [9]. Suppose that in order to determine the semantic similarity between words W1,W2, we can first find those subconcepts (or meanings) that contain words to be compared in the semantic network. In this case, the semantic similarity between the two can be represented by the shortest path between the two.

4. Research on big data computing model based on semantic data analysis technology

Big data's basic processing flow is information retrieval results, users generally prefer to see the results with the highest matching degree and the closest to the search conditions listed in the high display. The more likely it is to replace each other in different contexts without changing the syntactic and semantic structure of the text.

The research of semantic data analysis is based on the set of indicators based on application data source. This set consists of three types of indicators: statistical indicator \{Istati, I = 1 \leq p\}, syntax indicator (ISYN1, 2) and semantic indicator (ISEM1,2). By using regular expressions to define a category Kati, it plays the role of checking the syntax and semantics of strings, so that Katint can be a set of categories. Then RE can be defined as a collection of \{categories, RE\}.

The physical security requirement of data storage will be higher and higher, so the multi-copy and disaster recovery mechanism of data will be more demanding [10]. Among them, 16 kinds of relations, such as upper and lower relations and attribute-host relationship and so on, are described.

Based on the real-time requirement of business, Spark and flow processing framework S4.Storm, which supports iterative computing with Storm, Cloudar Impala, supporting online processing, is a distributed and fault-tolerant real-time computing system developed by BackType and later captured by Twitter. Storm belongs to flow processing platform, which is mostly used for real-time computing and updating database, as is shown by equation (3).

\[ W_f(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(x)\psi \left( \frac{x-b}{a} \right) dx \] (3)

Where is in this equation (3), \( W \) is Based on cloud computing architecture of big data, data storage and operation are provided in the form of services. Frequent use of metadata semantic analysis is not enough for us to understand the true meaning of big data correctly.

5. Conclusion

Big data computing model, that is, according to big data's the diverse data computing problems and requirements extracted and established a variety of high-level abstract (abstraction) or model (model). Our method is to propose a semantic data analysis to get a better understanding of the definition of data, and to improve the detection and correction of error.
Acknowledgments
This paper is supported by Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce, and also supported by the science and technology research major project of Henan province Education Department (17B520026).

References
[1] Aïcha Ben Salem1,2, Faouzi Boufares1, Sebastiao Correia2. Semantic Recognition of a Data Structure in Big-Data. Laboratory LIPN-UMR 7030-CNRS, University Paris 13. 2014.
[2] Meng Xiaofeng, kind. Big data Management: concept, Technology and Challenge. Computer Research and Development, 2013, 50 (1): 146 -169.
[3] Bell G, Hey T, Szalay A. Beyond the data deluge. Science, 2009, 323(5919): 1297-1298.
[4] Wei Lai. Research on folksonomy semantic Association recognition method based on online thesaurus. Office of Library and Information Services. 2011.
[5] Wang Shan, Wang Huiju, Qin Xiong pai, Zhou Shu. Architecture big data: challenges, status quo and prospects. Journal of computer Science, 2012, 34 (10): 1741-1752.
[6] Diao Qian, Zhang Huihui. Word weight and classification algorithm in automatic text classification. Journal of Chinese Information, 2000 (3): 2529.
[7] Le Xiaoqiu, Yang Chongjun. Recognition method of deep spatial semantics in unrestricted text. Computer engineering. 2010.
[8] Li Wenbo, Sun Le, Zhang Dakun. A new text classification algorithm based on Labeled-LDA model. Journal of computer Science, 2008 (4): 620-627.
[9] Hu Guangyong. Research on Data Security Storage Strategy Based on Cloud Computing. Computer Measurement and Control,2011,19(10)2539-2541.
[10] Li YL, Dong J. Study and Improvement of MapReduce based on Hadoop. Computer Engineering and Design. 2012, 33(8):3110-3116.