Novel Data Structure Modeling Model based on Hierarchical Data Analysis and Parsing

Chang Liu¹, Ting Han*, Changli Sun

Abstract

In this article, we describe the characteristics of large-scale modeling of the theme text of this site data and important progress in recent years. Topic modeling approach has attracted wide interest in the world, and promote a number of important data mining, development of computer vision and computational biology applications, including automatic text summaries, information retrieval, information recommendation, topic detection and tracking, natural scene understanding human action recognition and gene expression analysis. The main features of the model and the corresponding theme paper focuses on the text of this site data. Data with dynamic, high-end, multi-channel and distributed structure and the structure of the model is only part of the theme before modeling. The paper discussed in the framework of the unity of the three-dimensional Markov model of four structural features of the text of this site data modeling, and analysis of distributed computing and word combination of three-dimensional modeling topics Markov model and type fuzzy systems the possibility of applications. In addition to structural modeling for this site text data, also we discuss some of the three-dimensional Markov model energy minimization of machine learning algorithms.

Keywords: Data, Structure modeling, Paradigm, Hierarchical, Data analysis.

Introduction

With the rapid development of the Internet and related technologies, information presentation explosive growth, increasingly large-scale electronic documents for automatic analysis needs. Effectively organize, manage, and understand large-scale summary document, in order to fully identify and document the potential use of knowledge, making it more convenient and accurate access to the required information. Data nodes are usually indicates a link between the texts or image node contains a reference to multiple types of hyperlinks, labels and co-text content similarity and the like. Network data is the application of statistical and machine learning is an important object of study, modeling and analysis of network data aggregation interdisciplinary efforts, in the natural sciences, social sciences and information science have broad application.

¹Mudanjiang Medical University Center of modern educational technology, Mudanjiang City, Heilongjiang Province, 157011
*Corresponding author: Ting Han biedene@163.com
Statistical analysis of large-scale text data this site will not only provide useful predictive model, while providing large amounts of data statistics portray. Prediction model can predict the network data nodes simultaneously content (such as a word of the text) as well as a link between the node and the node (for example, a reference to a link between relationships and friendships between pages in the social network, scientific papers, etc.). Statistics can reveal hidden portrayed in the following network data clustering or community structure, thus helping to identify potential level of knowledge.

Topic model is a new statistical model, which uses a fast unsupervised machine learning algorithms to automatically organize, understand and summary of the main text, is a new method to solve such problems. Theme is a model for the analysis of large-scale text hierarchical Bayesian model, which uses a fast unsupervised machine learning algorithms to extract the low-dimensional mathematical representation, and thus the text from a word clustering high-dimensional sparse data word, mining is one of the hot areas of concern of the current machine learning and data.

Precision topic model can automatically discover and predict potentially important research direction, saving researchers valuable time and preliminary research funding, will play a major role in the biological, physical and chemical research from a number of scientific literature. This paper will be based on analysis of the text Ben data the main characteristics, the development of the current topic model and machine learning strategies are reviewed and discussed the design of new topic model and the development of appropriate new ideas fast machine learning algorithms, in order to characterize and use text network data-rich structural features to further improve the accuracy and stability of the themes modeling.

Figure 1. Hierarchical data analysis.
The Proposed Methodology

Data collection. Data collection refers to all types of structures such as the use of sensors, social networking and mobile Internet obtained, semi-structured and unstructured mass of data, which is the basis of all data analysis. Data collection to be resolved collect reliable data distributed high-speed, full-speed image data and other data collection techniques. Also design quality evaluation model developed data quality techniques and the data collection is generally divided into the large data IntelliSense Layer: includes data sensing system, network communication systems, sensor adapter system, identification system and access system hardware and software resources to realize the massive data intelligent identification, location tracking, access, transmission, signal conversion, monitoring, preliminary processing and management.

Data preprocessing. The process of data collection itself will be a lot of database, but if you want to achieve the purpose of effectively analyze vast amounts of data, it will import the data from the front end to a centralized large-scale distributed database, or distributed storage cluster, and in the lead We do discriminate on the basis of some simple, extraction, cleaning and other operations.

Storage and management of data. When we completed the acquisition of data, the need to store up a unified management, the main way is to establish the appropriate database, for unified management and calls. On this basis, the need to address big data can be stored and can be expressed, can be processed, several key issues and effective transmission reliability. The need to develop reliable distributed file system (DFS), to optimize the energy efficiency of storage, computing into the store, to redundant data and a cost-effective large data storage technology; and large distributed non-relational data management and processing technology, heterogeneous data, data fusion, data organization technology, research and large data modeling techniques, indexing, moving, backup, replication, visualization techniques.

Statistical analysis of the data. Under normal circumstances, the use of statistics and analysis is mainly distributed database, or distributed computing cluster to perform a normal analysis and subtotals for mass data storage therein so as to meet the most common analysis needs, in this regard, Some will use the real-time requirements of EMC GreenPlum, Oracle's Exadata, and MySQL-based storage Infobright column-like, while some batch or semi-structured data based on demand can use Hadoop. Statistics and Analysis of the main characteristics and challenges of this part is to analyze the large amount of data involved, the use of system resources, especially I/O will have a great occupation.
Data analysis and mining. The so-called data mining is the process of information from the database of a large number of incomplete, noisy, fuzzy, random data in practical application, revealing implicit, previously unknown and potentially valuable. Previous statistical analysis and process different is that data mining is generally not a pre-designed theme, mainly based on existing data above calculation algorithms, which play a prediction (Predict) effect, in order to achieve some of the high-level data analysis needs. The typical algorithm K-means clustering is used for statistical learning of SVM and a classification NaiveBayes, tools are mainly used for Hadoop Mahout like. Characteristics and challenges of the process is mainly used for mining algorithm is very complicated, and the amount of data and the calculation of the amount of calculation involved are large, common data mining algorithms are in the main single-threaded.

Multi-level data analysis methods. Standardized report is the first level of data analysis, requirements are relatively low, mainly through the appropriate statistical tools for data summarized obtain standardized report contains the main parameters. Similar to a sales enterprise financial statement monthly or quarterly.

Ad Hoc Reports. Users can own needs, the flexibility to select search criteria, the system can choose to generate the appropriate statistical reports based on user demand. Ad hoc queries and general application queries the biggest difference is the application of common query is custom development, and Ad Hoc queries all search criteria are defined by the user. Oriented high-level data analysis software, users are free to add indicators button you want to query together with the corresponding restrictions, you can generate visualize statistical results immediately, not only at a glance, and no difficulty operating.

Query Drilldown. Multidimensional analysis refers collectively, methods commonly used means of visual analysis of the data model with multiple dimensions and metrics will be composed include: drill, volume, sliced (cut), rotation, and other analysis operations. In order to analyze the data, make the analysis, decision makers from multiple angles more sides of observational data, giving insight into the information contained in the data and content.
Volume on the implementation of operations in the data cube gathered by the rise in the level or dimension by eliminating one or more generally to observe certain dimensional data and another case on the volume by eliminating one or more dimensions to observe more general data. Drill a decline in the level of peacekeeping or observation data by introducing a certain dimension or in more detail. Slice in a cube dimension given data selection operation, the result is to get a slice of a two-dimensional plane data (selection operation is cut in two or more dimensions given data cube, the result of the cuts is to get a sub-cube). The shaft is relatively simple, is to change the dimension direction.

Statistically Analysis. We know that probability theory is the basis of mathematical statistics; mathematical statistics are random variables in their research on the basis of probability theory and application of knowledge to make a reasonable estimate, inferred and predictions. On various probability distributions are discussed in mathematical statistics as a statistical model to analyze process data with random errors. Typical mathematical statistics method includes parameter estimation, hypothesis testing and regression analysis. The statistical analysis is mainly on issues of concern to users analytical methods to infer, predict and control.

Alerts. Dashboard for monitoring some key indicators. Simulation analysis is to be dynamically adjusted by the operator's control (such as a slider, adjustable knob, selection box, etc.) to control some of the parameters of the model behavior management decisions. When the operator through the control panel in the model parameter values or variable values adjusted curve elements, cylindrical or group analysis indicators drawing occurs corresponding movement, and this movement reflects that change the parameter impact on the behavior of the model, if such change causes a change in the model the optimal solution or other key figures, it is possible at any time on the conclusions of this change is displayed correctly.

Forecasting. In the statistical analysis and data mining, predictions of the future have a lot of mathematical models and associated algorithms for solving specific problems. Its core idea is to find the development pattern data from historical data, then these patterns as a fulcrum, can predict the future.

Predictive Modeling. As data analysts continue to explore data mining technology, there have been many forecasting models and with the corresponding algorithm, but it is difficult to determine if a model is the most accurate, because different areas, different conditions, the corresponding prediction model is not the same, so there is no unified optimal model, there is only selective optimal model.

Optimization. Because the optimization problem can often bring huge gains, through a series of practical optimization, can be significantly improved earnings. The so-called optimization is to select the best solution from a finite or infinite number of possible scenarios. If you can simply judge, we can determine the optimal solution that is the best. But the truth is not so simple, so optimization techniques have been developed a series of theories to solve practical problems.
Conclusion

Multi-level data analysis is very important in the management, because the value it produces is entirely on the level of real data, while an enterprise, improving data collection mode is to improve the management process, enterprises in line with the times and standardize the management process is essential. Multi-level data analysis can promptly correct erroneous business decisions, it is possible to start real-time tracking of progress, but also to understand the situation and cost control personnel ideological dynamics.

Acknowledgement

This paper is supported by the Mudanjiang Medical College Education and the Teaching Reform Research Project: Research and Practice of the Mixed Teaching Mode Based on the Mutual Class (NO. JY2016051).

References


