Creative Education, 2016, 7, 2868-2875
http://www.scirp.org/journal/ce
ISSN Online: 2151-4771
ISSN Print: 2151-4755

# Education Web Information Retrieval and Classification with Big Data Analysis

**Rurui Zhou**

Office of Educational Administration, Yangtze University, Jingzhou, China
Email: 68681700@qq.com

**How to cite this paper:** Zhou, R. R. (2016). Education Web Information Retrieval and Classification with Big Data Analysis. *Creative Education, 7,* 2868-2875.
http://dx.doi.org/10.4236/ce.2016.718265

**Received:** November 28, 2016
**Accepted:** December 26, 2016
**Published:** December 29, 2016



## Abstract

This paper discusses the text mining method to obtain the education web information rules. The method can be applied to words and parts of speech in texts to record the two-tier structure of automatic mining. On the basis of the initial discovery of the labeling rules, we put forward the linguistic features based on words to expand the application of the approaches. The web information retrieval mechanism is proposed based on web content segmentation. In addition, we build a rule matching method to improve performance of rule utilization, in terms of information of interest to the user and the extent to the interesting part. In conclusion, the use of automatic labeling rules can make part of the tagging accuracy rate reach a new height.

## Keywords

Education Informatization, Big Data, Text Mining, Information Retrieval, Text Categorization, Web Categorization, Web Page Segmentation, Concept Semantic Space

## 1. Introduction

Recently, the education information on the Internet is getting more and more abundant, more and more extensive; it has become the most important information for multiple data source (Wu et al., 2014; Liu & Motoda, 2012). To help users quickly and accurately find and classify useful information online, in terms of a wide range of application background and practical value, education web information with big data analysis has become a research hotspot (Fan & Bifet, 2013; Baradwaj & Pal, 2011). Education managers, on the one hand are to use big data mining technology to improve retrieval and classification accuracy. On the other hand, through the research on these issues, they study the knowledge representation of Internet information, in light of similarity

measurement, large-scale big data mining, the effective use of massive information, and retrieval and classification methods, to do some meaningful exploration (Mukhopadhyay et al., 2014; Tsytsarau & Palpanas, 2012).

Also, the web information retrieval method can be built based on web content segmentation. The method is based on semi-structured features of web pages point, in accordance with the XML tags and the contents of the page will be page segmentation. In the establishment of XML tag, we can establish tree on the basis of education application. According to the user's query, we can make full use of regional information to sort the relevant search results. The experimental results show that the method can improve the accuracy of the search engine, and the design of the next generation of search engines to provide a useful reference (Dumais, 2016; Ghorab et al., 2013).

In this paper, the concept of conceptual semantic space and its similarity measure based on this space can be developed. In concept, conceptual relevance and semantic dictionary, with the concept semantic space, we propose a new similarity measure method. The meanings of education linguistic entities are established via a "meeting of minds". The concepts in the minds of communicating individuals (Campos, 2015; Ting et al., 2013) are modeled as convex regions in conceptual spaces.

The paper is organized as follows: the first section is the introduction. The second section is about education text and web mining. The third section addresses education text mining. The fourth section draws a conclusion.

## 2. Education Text and Web Mining

### 2.1. Education Big Data Mining Processes

Big data with data mining (Goeuriot et al., 2013; More, 2015; Shen et al., 2014) is a database knowledge discovery of education information, which is specifically responsible for discovering knowledge. This process is an interactive, cycle of the overall process, thereby identifying effective, novel, potentially useful, and most useful big data from a big data set. The final understandable pattern of nontrivial processes is shown for data analysis. Although education information includes data mining in addition to data preparation, the discovery, interpretation and many other links should be developed. Because of the important role of big data mining in education information, the data mining process can be considered big data mining to contain the education information in the many links.

Big data mining is for the large, incomplete, noisy, fuzzy, random big data set knowledge, for non-trivial processes that are not valid, novel, potentially useful, and ultimately understandable. It is a wide range of interdisciplinary subjects, including big data, machine learning, mathematical statistics, neural networks, pattern recognition, rough set, fuzzy mathematics and other related technologies. The interpretation and evaluation of the results is shown in **Figure 1**.

Consequently, the preparation for education information analysis can be divided into three parts: *big data selection*, *big data pre-processing*, and *big data transformation*:

- The goal of big data selection is to determine the target of the discovery task, for the target big data.

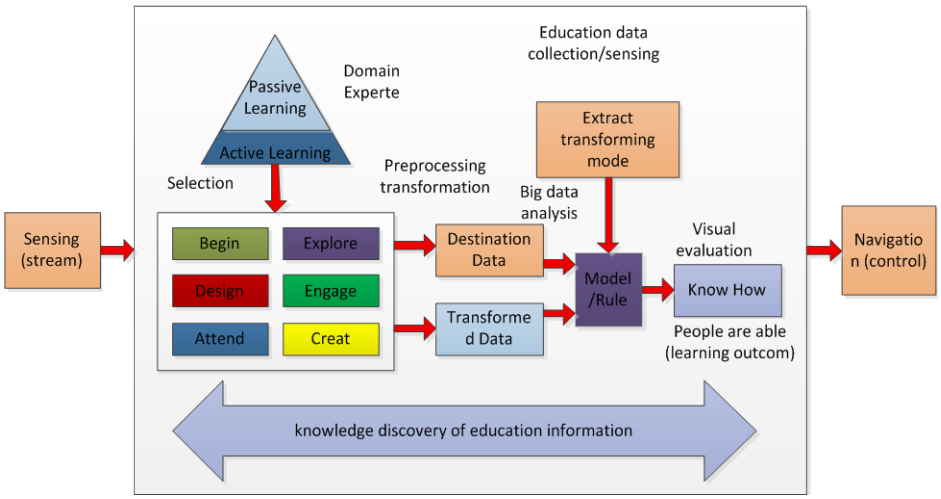


**Figure 1.** The process of education data mining.

- In general, big data preprocessing may include eliminating noise, deduce the calculation of missing big data, eliminate duplication of records and complete big data type conversion, such as the number of consecutive values. According to the conversion of discrete big data, the discrete type conversion has continuous value type in order to facilitate the symbolic induction.
- The main purpose of big data transformation is to reduce the big data dimension or dimensionality reduction, i.e., from the initial feature to find truly useful features to reduce big data mining number of features or variables.

Big data mining stage to determine the first task or what is the purpose of mining, such as big data summary, classification, clustering, association rule discovery or sequence pattern discovery. The same task can be achieved with different methods. There are two options to achieve the method considerations. Firstly, different big data have different characteristics, and therefore need to be related to the method to extract. Secondly, for the user or the actual operation of the system requirements, some users may wish to obtain descriptive, easy to understand the knowledge, while the purpose of some users or systems is to obtain predictive accuracy as high as possible.

The patterns discovered during the big data mining phase may be subject to redundancy by user or machine evaluation. The entire discovery process has to return to the discovery phase before, such as re-select the big data, using the new big data transformation method. We set new big data mining parameter values, and even for an excavation method, such as when the task is found classification. There are a variety of classification methods, different methods have different effects on different big data. In addition, the mining results by the end are for the human user, they may be necessary to visualize the pattern of discovery, or the results for another user's understandable representation.

## 2.2. Classification of Education Texting Data

Researchers have done a lot of research on the mining of association rules. The work

includes the optimization of the original method, such as the introduction of random sampling, the idea of parallel to improve. The efficiency of method mining rules is required for the application of association rules, or the mining association rules for effective organization and interpretation.

Big data classification is based on the classification model in accordance with the attribute value of the big data collection classification. It is one of the big data mining problems in the process. The goal is to extract classification rules. Classification is a mentor to learn, generally need to have one. Training samples big data set is as inputs. A training set consists of a set of big database records or mechanism, and each mechanism is an eigenvector consisting of the values of the fields.

Also, clustering can be called unsupervised classification without training set required. Clustering is a group of individuals in accordance with the phase. Its purpose is to make individuals belonging to the same category. More specifically, clustering methods include statistical methods, machine learning methods, spatial big database methods, big data mining methods, and so on.

Moreover, visualization is the process of transforming big data, information, and knowledge into visual representations. Visualization technology provides an interface between the two most powerful information processing systems, the human and the computer. Using effective visual interface, we can quickly and efficiently deal with large amounts of big data to find hidden features, relationships, patterns and trends. Further, visualization has a wide and important field, which can lead to new insights. These works involve the visual search of text.

It is necessary to categorize these big data mining methods. Being describe or describe an operator law involves three parts: *input*, *output* and *processing*. The input is the big data mining method with a variety of shapes. The output of the method is the knowledge or pattern to be discovered. Also, the processing of the method involves a specific search method. From the input, output and processing of the method three points, we can determine such a few species classification criteria, i.e., mining objects, mining tasks, and mining methods. This classification refers to a general process related to categorization, in which ideas and objects are recognized, differentiated, and understood.

According to the mining task, there are the following types of knowledge discovery tasks: classification or prediction model knowledge discovery, big data summarization, big data clustering, association rule discovery, sequence pattern discovery, dependency or dependency models anomalies and trend finding, among others.

## 3. Education Text Mining

At present, the research of text big data mining is in the early stage of development, for its meaning, process, function, etc. There is no uniform conclusion. Different researchers have a different understanding of the study of text mining. Also we have the text data mining, roughly equivalent to text analytics, in terms of the process of deriving high-quality information from text. Generally, text mining is considered to be in a large

number of text collections or corpus, which implied that people are interested in useful patterns and knowledge. Most of the knowledge is found in the big database function, such as dependency analysis, classification, clustering, deviation detection, etc. They can be in the text mining or may be implemented. Obviously, this definition will be text mining as a big data mining from structured big database to no knot. Also, text mining is a new research area, through the use of big data mining, machine learning, natural language processing, and information retrieval. It involves preprocessing, intermediate forms of document collections.

Text mining is another integrated technology, involving big data mining, natural language processing, computational linguistics, information retrieval and classification, knowledge management and other fields. The object of text mining is natural language, so it inevitably involves knowledge in the field of linguistics.

### 3.1. Education Text Extraction

These mainly involve the extraction of phrases in the text, such as mining phrases, finding co-occurrence phrases, discovery of maximum frequency phrases, and so on. Many of this work involve association rule techniques in big data mining. There are four steps to the text mining framework:

1) *Information Retrieval*: The main goal is to retrieve the collection of documents related to the task.

2) *Information Extraction*: From the selected collection of documents to extract information, a typical way is to fill with user-specified desired information template. It is the essence of the formation of the big database.

3) *Information Mining*: The use of standard big data mining technology to find patterns from the big data.

4) *Explanation*: The best way to interpret the patterns found during the extraction phase is by nature. The framework is mainly from the perspective of big data mining to study the text mining, the lack of text mining. The use of language information processing technology is with the mining process.

The use of unsupervised neural network method is to automatically organize the document collection two-dimensional view. More views can provide a more comprehensive picture of the set of unknown documents. Among these environment, phase similar documents appear in the adjacent view area, each area contains some description of the keywords, thereby helping users familiar with some unknown areas. We emphasize the visualization while improving the results of the information retrieval using document views. The framework of education text mining is shown in **Figure 2**.

Also, text mining is an emerging technology for analyzing unstructured document sets whose goals, according to the extraction of interesting, complex patterns and knowledge. Clearly, text mining is an interdisciplinary subject involving letters Information retrieval, text understanding, information extraction, clustering, classification, visualization, big database technology, machine learning and big data mining.

In the framework of text mining, text mining can be divided into two stages: *text*

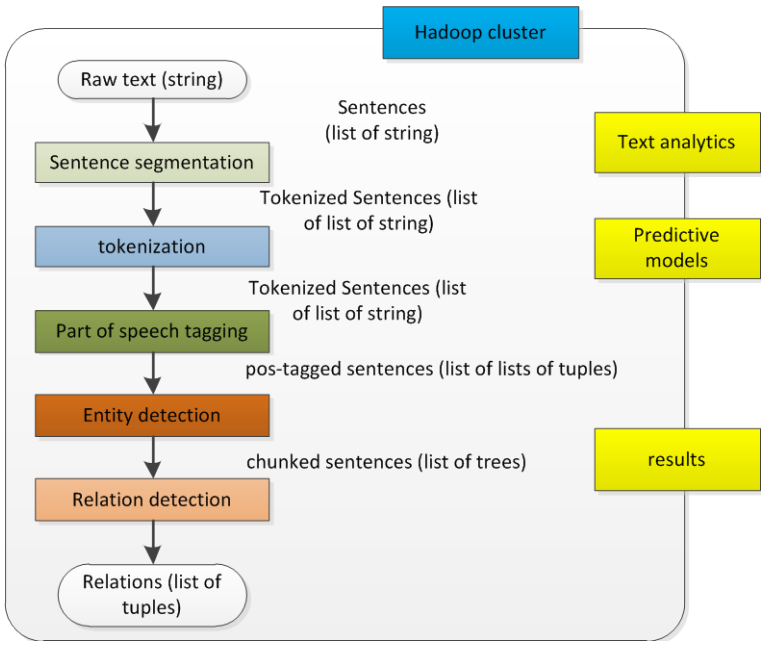


**Figure 2.** The framework of educational text mining.

*refinement refining* and *knowledge distillation*. Text refinement will be free when the knowledge is extracted from the intermediate form to infer the patterns and knowledge. Intermediate forms can be semi-structured forms, such as concept maps or structured relational tables. In general, the concept is based on the intermediate form depending on the domain, and the document-based intermediate forms can be domain-independent. In contrast, the e-learning course can provide basic underpinning knowledge on distillation and its application. The form can be transformed into the relevant objects of interest in a particular domain formula. Document-based intermediate forms can be obtained between the mining mode and the relationship between documents, and thus having document classification, clustering, and visualization. The mining of concept-based intermediate forms is available objects, patterns and relationships between concepts. Thus, prediction models and associations fall into this category.

Feature extraction for shallow analysis is to further processing for feature selection, the use of heuristic information and part of speech information. Mining results have the various types of documents in the feature words, names, place names, organization name, compound word, acronym, relationship, amount, date, and language identification. It is applied to the analysis of customer feedback information.

Classification of the first feature selection is in the training phase to generate a special index, mainly through the search each category of keywords in the rules of induction. In the identification phase of the new document sentenced, the result returns a set of class labels and the corresponding confidence level. Finally, we can organize our intranet documents into folders.

Below we list text mining products and applications. The products can be divided into two categories: one group is text analysis, information retrieval, information ex-

traction, classification and abstracts, the other group is the organization of the document, visualization and navigation.

A feature set is established for each ambiguous word. Feature set contains morphological features (singular and plural), words characteristics and word collocation. The observations in the big dataset are then clustered. The semantic discrepancies are then used to obtain the patterns. Web mining is the use of big data mining technology from web documents and services in the automatic discovery and access to information. In general, web mining can be considered packages including the following four tasks:

- *Resource discovery*: Tasks associated with web document retrieval.
- *Information selection and preprocessing*: Automatically selecting and preprocessing specific information in a web resource.
- *Generalization*: Discovering the general pattern of single or multiple Web sites.
- *Analysis*: Verification and interpretation of the extracted patterns.

Resource discovery is a process of retrieving big data from online resources on the Web, including when newsletter, news, and text content that removes XML tags. Information selection and preprocessing is an information transfer process. It can remove stop words, stemming, find phrases in the training set, and get it relationship or logical representation. The generalization process is the application of big data mining technology to obtain knowledge. The final analysis is the verification and interpretation of the mining results can play a role in this process.

### 3.2. Web Text Information

Also, education web pages contain text information. In this sense, the text mining is actually a web mining for using the web document within the purepart. The structure of the text and its XML files, part of the work are combined with the hyperlink structure between documents. One of the most basic problems in web content mining is the representation of text or web pages. Thus, content extracts many topics, such as classification, clustering, learning the relationship between pages, rules or patterns of extraction and so on with this phase turn off.

Classical text representations use vector space representations. This method is used in the training set word as a feature, while ignoring the order of occurrence of the word and a single word statistics. The value of a feature can be a Boolean value, or based on the frequency value, such as the word frequency multiplied by the frequency of the inverted document. At the same time, we have the feature selection, such as remove disable word, and some machine learning techniques, such as information gain, mutual information, entropy and probability ratio. In addition, the use of semantic index can be achieved vector drop dimension.

## 4. Conclusion

This paper proposes a dynamic interest learning method, where the user does not directly edit the interest description file. The method only uses less human-computer interaction, which can classify keywords and calculate the degree of interest of the user, in

order to obtain the initial personalized interest. A description file is used as a basis for interest identification. Through this document, it can be used to determine whether some of the literatures of the user are interested. In order to achieve an effective personalized service, we use a local autonomous agent to perceive the user's behavior and monitor the user's interest in real time. These behaviors include dwell time when the user accesses. Finally, the number of visits, saving, edition, modification and other actions can be obtained, while the user inputs keywords as dynamic update interest file to consider.

## References


Baradwaj, B. K., & Pal, S. (2011). Mining Educational Big Data to Analyze Students' Performance. *International Journal of Advanced Computer Science and Applications, 2,* 63-69.

Campos, R., Dias, G., Jorge, A. M. et al. (2015). Survey of Temporal Information Retrieval and Related Applications. *ACM Computing Surveys (CSUR), 47,* 15.

Dumais, S., Cutrell, E., Cadiz, J. J. et al. (2016). Stuff I've Seen: A System for Personal Information Retrieval and Re-Use. *ACM SIGIR Forum, 49,* 28-35. https://doi.org/10.1145/2888422.2888425

Fan, W., & Bifet, A. (2013). Mining Big Data: Current Status, and Forecast to the Future. *ACM sIGKDD Explorations Newsletter, 14,* 1-5. https://doi.org/10.1145/2481244.2481246

Ghorab, M. R., Zhou, D., O'Connor, A. et al. (2013). Personalised Information Retrieval: Survey and Classification. *User Modeling and User-Adapted Interaction, 23,* 381-443. https://doi.org/10.1007/s11257-012-9124-1

Goeuriot, L., Jones, G. J. F., Kelly, L. et al. (2013). ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information Retrieval to Address Patients' Questions When Reading Clinical Reports. *CLEF 201*3 Online Working Notes, 8138.

Liu, H., & Motoda, H. (2012). *Feature Selection for Knowledge Discovery and Big Data Mining.* Berlin: Springer Science & Business Media.

Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., et al. (2014). A Survey of Multiobjective Evolutionary Methods for Data Mining: Part I. *IEEE Transactions on Evolutionary Computation, 18,* 4-19. https://doi.org/10.1109/TEVC.2013.2290086

More, L. (2015). Analytical Study of Information Retrieval Techniques and Modified Model of Search Engine. *International Journal on Computer Science and Engineering, 7,* 57.

Shen, Y., He, X., Gao, J. et al. (2014). Learning Semantic Representations Using Convolutional Neural Networks for Web Search (pp. 373-374). *Proceedings of the 23rd International Conference on World Wide Web*, Seoul, 7-11 April 2014. https://doi.org/10.1145/2567948.2577348

Tsytsarau, M., & Palpanas, T. (2012). Survey on Mining Subjective Big Data on the Web. *Big Data Mining and Knowledge Discovery, 24,* 478-514. https://doi.org/10.1007/s10618-011-0238-6

Ting, S. L., See-To, E. W. K., & Tse, Y. K. (2013). Web Information Retrieval for Health Professionals. J*ournal of Medical Systems, 37,* 1-14. https://doi.org/10.1007/s10916-013-9946-3

Wu, X., Zhu, X., Wu, G. Q. et al. (2014). Big Data Mining with Big Bigdata. *IEEE Transactions on Knowledge and Big Data Engineering, 26,* 97-107. https://doi.org/10.1109/TKDE.2013.109