Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(6):2198-2202

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Analysis on the technology improvement of the library network information retrieval efficiency

He Lingling[1] and Liu Yiwei[2]

[1]Library, National University of Defense Technology, Changsha, Hunan, China
[2]Department of Basic Education, Military Economics Academy, Wuhan, Hubei, China

---

ABSTRACT

This article provides a viable method to measure the barriers in Air Transport Services Trade and the economic effects of the liberalization, which is an initial attempt in China Air Transport Services Research. Through the research, this paper concludes that opening transport services not only makes a great contribution to the services trade, but also plays a significant role in promoting goods trade and economic growth. Opening transport services could reduce the cost of transport services, and it could have the same effect as reducing the protective tariff. One effect of reducing the cost is the increase of goods trade volume. Another conclusion of this paper is about the economic tie between the cost of transport services and economic growth, that is, if the transport costs double, the annual economic growth rate would be halved.

Key words: Air transport services; Trade barriers; Liberalization; Economic effect

---

INTRODUCTION

In the modern information environment, library emphasis increasingly on the development and utilization of the network information resources, being discrete, disorder, diversity, the vast online information inconvenience people in the process of retrieving information. So how to search the Internet information needed efficiently is a big problem. An excellent computer information retrieval system can improve the accuracy and convenience greatly. Compared with manual search, the biggest advantages of computer retrieval system lie in that it can classify the information automatically, form the indexing and database in the vast information resources and search the accurate and comprehensive information rapidly. So, the key to the performance of retrieval is of indexing of huge amounts of information.

1. The Characteristics of the Current Library Network Information Resources

It consists of various data sources (e.g., news articles, research papers, books, digital library and Web page) of a large number of documents. Because of the rapid growth of information available electronic formats, such as electronic publications, electronic mail, CD-ram, and the world wide web(it can be also seen as a huge, interconnected, dynamic database), etc., most documents stored in the database data is so-called semi-structured data (semistructure data), which is neither structureless nor structural completely .For example, a document may contain structure fields such as the title, author, publication date, length, classification, and so on, it may also contain a lot of non-result text element, like abstract and content. In the late database research field, there are a large number of research of relevant semi-structured data modeling and implementation. what's more, information retrieval technology, such as text mining, has been used to deal with unstructured documents. [1]

2. The Key of Library Network Information Retrieval Technology, Automatic Classification

2.1 Text Mining and Automatic Classification

The traditional information retrieval technology has not adapted to the needs of increasing amounts of text data

processing. Traditional taxonomy for formal publication of the printed literatures, attaches great importance to the integrity, the logic of knowledge system, on the basis of subject classification and knowledge classification, category structure is rigorous, has adopted the alternant columns, compiling annotation, equipping, and see to strengthen the connection between the category, Traditional taxonomy can not adjust to the diversified and highly dynamic disorder online information resources .A typical traditional retrieval is to find out a small part related to an individual or a user from a large number of documents. It is difficult to form effective query, analyze and extract useful information from data without knowing the contents of the document .Users need related tools to complete the comparison of different documents, as well as the arrangement of important and relevant documents, or to find out the patterns or trends of multi-documents. [2]

The main task of the information retrieval is to provide users with the corresponding document information according to their request of the query. With the increasing volume of information processing and the demand for information, an analysis and processing of the level of document database will become the inexorable developing trend. As an effective way to acquire knowledge, text mining is a text information processing technology which developed on the basis of in information retrieval .It is one of the emerging branch of knowledge of management research field and provides an effective ways for the text information collection, analysis and mining .The traditional information retrieval or access to information, mainly retrieve the relevant document information from the document database according to the query conditions provided by the customers. In order to improve the accuracy of the information access, retrieval system increases the related processing, such as the methods of document classification, automatic abstract, and keywords automatic extract, users can find the needed information easily.

Automatic document classification is an important text mining work. Because of numerous online documentation, classification and organization for it automatically for the convince of retrieval and analysis of the document is of great importance. Automatic classification of network information is to collect and analyze of classified object and include it into relevant category of certain classified system with the assistance or the replacement by computer .On the basis of the principle of co-occurrence, Automatic classification analysis statistically by extracting the content features of network information, identify the words representing its information content furthest, and then set similarity analysis of the classified system words , determine the type of the information belongs to, entitle them with certain classified logo(verbal, class number and a sort of code).

2.2 The Types of Automatic Classification

Text categorization refers to the computer (or other entities or objects) classify and mark the text according to a certain classification system or standard.

There is no essential difference between the problem of text categorization and other classifications. The method put in a nutshell in matching which is unlikely to be complete according to some of the characteristics of classified data of course. So the optimal matching results must be chosen to finish the classification accordingly. [3]

1. the Word Matching Method

Word matching method is the earliest proposed classification algorithms. It only judges if the document belongs to a certain category according to whether it include the same words with taxon (plus the processing with the synonymous at most) .Obviously, this mechanical method is too simple to bring good classification effect.

2. Knowledge Engineering Method

Knowledge engineering method defines a large number of inference rules for each category with the help of professional. if a document can meet these inference rules, then it can be predicated belongs to the category. The degree of match here with specific rules become the characteristic of the text. Due to the artificial predicating factors in the system, the degree of accuracy is greatly improved than word matching method .But the downside of this approach is still obvious, for example, the quality of classification relies on these rules severely, that is to say, it relies on "people" who setting the rules.

3. Statistical Learning Method

Statistical learning method requires a number of accurate classified documents by artificial document as study materials, computer unearthed some effective classified rules from the documents, this process is called training, and the summarized collection of rules is often referred to as classifier. After the training, it is necessary for the classifier to categorize the documents the computer had never seen before.

Nowadays, statistical learning methods have become the absolute mainstream in text categorization. The main reason is that many of these technologies have a solid theoretical foundation, clear evaluation criteria, as well as the actual good performance.

The frequently-used classification algorithm consists of following methods: the decision tree, Rocchio, naive Bayesian and neural network, support vector machine (SVM), linear least square fitting, KNN, genetic algorithm, the maximum entropy, Generalized the Instance Set, etc.

Rocchio algorithm: It is called the "centroid" to a take a mean to get a new vector in the sample document of a category called the "centroid", which is the most representative vector. Compare the degree of similarity between the new document and centroid to confirm if it does belong to this category.

Naive Bayes algorithm: It concerns about the probability of a category that the document belongs to, which is equal to the comprehensive expression of probability of the category of each word of the document. And the probability of the category that each word belongs to can be roughly estimated through its frequency in the category training document to a certain extent, thus makes it feasible for the whole calculation process.

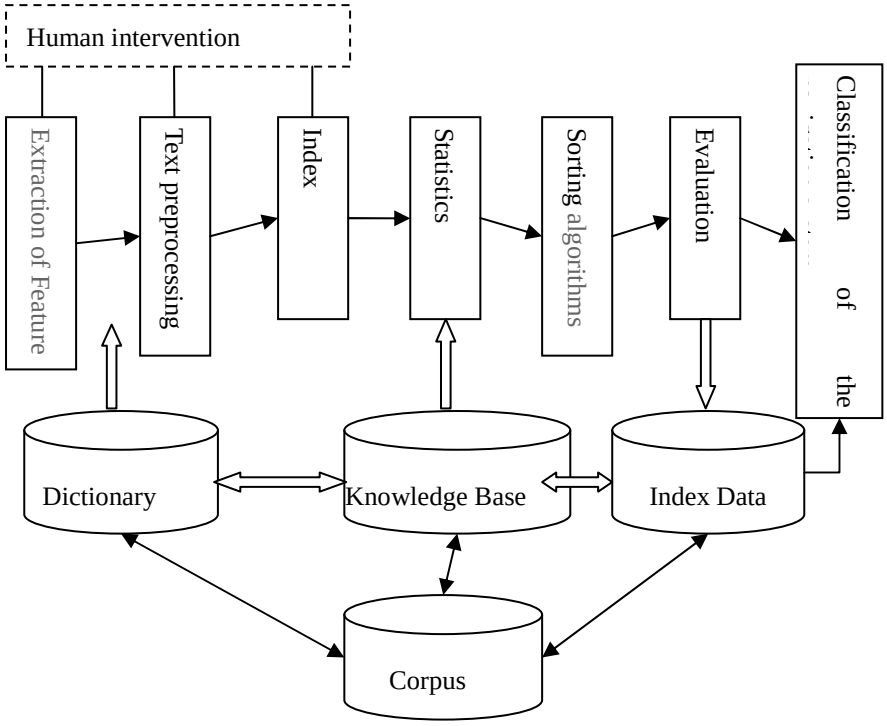
Nearest neighbor (KNN) algorithm: calculate the similarity of the feature vector of the new document and the vector of training document in each document after the new document is given and get the K most similar documents with the nearest article and predicate the category of the new document belongs to according to the category of the K documents, which is very suitable for the demand of the variational classification criteria.

SVM (Support Vector Machine): Based on VC dimension theory and structure risk minimum principle of statistical learning theory, this method seeks the best compromise between the complexity of the model (i.e. the specific learning Accuracy of training samples, Accuracy) and learning ability (i.e., not wrongly's ability to identify random sample) to get the best generalization ability (or the generalization ability) according to the limited sample information.

2.3 The General Steps of the Automatic Categorization for the Network Information

Document classification includes the processes of text expression which can be subdivided into the processes of text preprocessing, indexing and statistics, feature extraction and so on, text categorization classifier and training, the evaluation and feedback of classification results. [4]

As shown in the figure below is the processor of Automatic classification of network information:



The overall function module for text classification system:

(1) Pretreatment: Format the raw corpus to unify the text format for the subsequent unification management. Text preprocessing includes the transformation of different text formats, the special position and tags in the text of the processing, the generation of textual XML representation and other steps.

(2) Index: the document is decomposed into the basic processing unit, and at the same time reduces the subsequent processing overhead. Indexes including the word item extraction, selection and weighted index entries, etc. Term extraction refers to the term split from the text to express the content; the selection of index entries depends on how to deal with the meaningful text unit and the combination of these units natural language rules.
(3) Statistics: word frequency statistics, the related probability of term (word, concept) and classification.
(4) Feature extraction: extracted from the document to reflect the characteristics of document theme, including feature vector, removing stop words and function words, the concept of network, etc. Words are weighted by analyzing the importance of it in the text and the frequency statistics and the weight determine which words can be used as the content feature of the subject content of the text.
(5) Classifier: Classifier data contains class center, words and the related probability of classification. First extract the feature and classification of the knowledge base to format the vector space model to conduct similarity matching to cover each feature item as the main category (including crossing category). Then find out the category matching with the other feature categories as the secondary category (including cross category), which is a critical step in automatic classification.
(6) Evaluation: classifier analysis of the test results. Work out the text abstract and other identifying through the artificial or computer.
(7) To store the text in the database after indexing and description.
(8) To classify the site in the navigation system and sort it automatically.

2.4 The Evaluation Methods of Automatic Text Classification

Assume that the artificial classification is entirely correct and eliminate factors of personal thinking differences, the closer the results is to artificial classification ones, the higher the accuracy of the classification is, which implied the two indicators of the evaluation text classification system, the precision and recall. The former is the ratio that all the judgmental texts is identical to the texts in the artificial classification results, its mathematical formula is as follows:

$$precision = \frac{\text{the correct number of the text classification}}{\text{the actual number of text classification}}$$

Recall is the ratio of the text which is consistent with the one in the artificial classification results; the mathematical formula is as follows:

$$recall = \frac{\text{the correct number of the text classification}}{\text{the proper number of text}},$$

Precision and recall reflects the two different aspects of quality, both of them must be considered comprehensively and cannot be ignored, therefore, there is a new evaluation index, the F1 test values, and its mathematical formula is as follows:

$$\text{F1 test values} = \frac{\text{accuracy rate} \times \text{recall ratio} \times 2}{\text{accuracy rate} + \text{recall ratio}}$$

In addition, there are two methods, micro average and macro average, to calculate the precision, recall and F1 value method.

Micro average: calculate each kind of precision, recall and F1 value.
Macro average: calculate the total precision, recall and F1 value.
The target of all the text classification system is to make its process more accurate and rapid.

3. The Development Trend of the Library Network Information Retrieval

3.1 Personalized Development Trend

On the one hand, the development of the network make it possible that the library database in the network environment develop in the direction of large-scaling and integration and makes a batch of even a computer terminal or a site can become a small database simultaneously. The data that is available to be checked provided by this small database check often is very professional and personalized, and with the continuous development of our country library network, the computer terminals and the site will be more and more, therefore, the development trend of personalized is one of the features of library information retrieval in the future.

3.2 Standardization Development Trend

The diversity and complexity, as well as the dispersion and disorder of the network information affect the library network information retrieval seriously. Therefore, the measurement standard of network information must be normalized as soon as possible to end the current state of disorder and a set of network information recording, data organization, information retrieval and retrieval results standardized criteria must be set up. Standardization is the priority of network information retrieval and also the development trend of information retrieval in the network environment.

CONCLUSION

The library network information resource development is rapid in our country at present, the trend of sharing and complementary of network resource between libraries is more and more obvious. Hence, the high efficiency of computer retrieval system must be studied to make it possible for the readers to retrieve the needed information in the boundless network resources while the automatic classification technology play a crucial role. With the improvement of retrieval efficiency, the library network information resources utilization will be promoted greatly which can provide the readers with more efficient and high quality service.

REFERENCES

[1]Wang Lin. *Library Science Research*, **2002**(2)

[2]Li Xiaoming, Yan Hongfei, Wang Jimin .*Search Engine- Principle, Technology and System.* [D]. Beijing：Science Press.**2004**

[3]Chen Shunian,. *Automatic Classification Problem of Network Information*[J]. *Library Journal*,**2001**(10)

[4]Su Weifeng, Li Shaozi . *Engineering and Application of Computer*, **2002**(6)