Research of Automatic Classification of Retrieval Results for Chinese Internet Fraud Information Management Platform

Hu Liang[1, a *], Zhu YuChi[1, b], Xu Jing[1, c]

[1]JiangXi Police College, Department of Humanities and Management, XingWan Road, XinJian County, NanChang City, JiangXi Province, P.R.China

[a]huliang_thu@163.com, [b]zhuyuchi_jxga@163.com, [c]xujing_jxga@163.com

Keywords: information retrieval, automatic classification, internet fraud, information management

Abstract. For solving some problems of information retrieval in the Chinese internet fraud information management platform, the automatic classification model of search results based on the personalized service was proposed. The model computes the K nearest neighbor of the search results in the training sample set, finds the class of the most of the K nearest neighbor, and then classifies the search results as the class, and divides the search results into a simple hierarchy, which can both convenient the user and improve search efficiency. The performance test presents that the model can improve the search accuracy and quality of the specified internet fraud information.

Introduction

The traditional information retrieval in the information management platform is that firstly query using the search keywords, and then sort the search results by relevance. But the problems of polysemy, multi-word synonymous and differences in the existence of user habits, make the search accuracy and quality couldn’t fully meet user needs, so how to build the search results displaying form which facilitate the user to quickly find the requirement is a very important research direction. In our literature review, statistical data show that the average search keyword is short in length and rarely use specialized query operators, thus we consider to utilize the phrase or location adjacent relationship between the search keywords for improving the retrieval results in the actual retrieval[1,2]. In order to promote the search quality, this study organizes the search results in a hierarchical manner, that users can easily select a category, increase the search efficiency and find research information faster.

Classification of the Chinese internet fraud information

In this study, the internet fraud information is defined as the primary data after once processing. It could identify the physical symbols to reflect the objective world, which includes documents, statements, figures and their data produced in research activities[3,4,5]. The concept includes the meaning of two aspects: the objectivity and identification of the data. For example, we regard the rice as data, then the flour is the primary data after once processing and the noodle is data after secondary processing. Here the boundaries of the primary data is the processing degrees. Our standard is that information is new information which is just simply classified or graded physically, has not put too much labor, does not exceed the value of its own value-added, has no chemical changes and no other added information.

The classification of the Chinese internet fraud information is a very complex issue[6,7]. In addition to the reasons of a wide variety and surprising number of information, the classification criteria is not unified and there exists many different habitual classification method. The internet fraud information classification in this research refers to the process of automatically discriminating information categories according to the content of the information process under a given classification system. The information is a multi-layer structure[8,9,10,11]. Such as the internet fraud information is the class for the first level; Source and type is the class for the second level; Source can be divided into news, police database and the third platform which is the third level; Type can also be divided into Micro-blog, WeChat, QQ, Email, Taobao, web site and online games, which is the third level. It can also be divided

© 2015. The authors - Published by Atlantis Press

continued. This is determined by the rich hierarchical nature of the objective world. The multi-layer classification for humankind is natural and easy to understand. However, it is quite complex for the computer. Complex classification will increase the computational complexity and reduce efficiency. This study will simplify the Chinese internet fraud information into to three-layer tree. The first layer is the major categories of information, the second layer is subclasses of the major class, such as source and type, the third layer is the theme of subclasses, such as the news, police database, etc. as shown in Fig. 1.

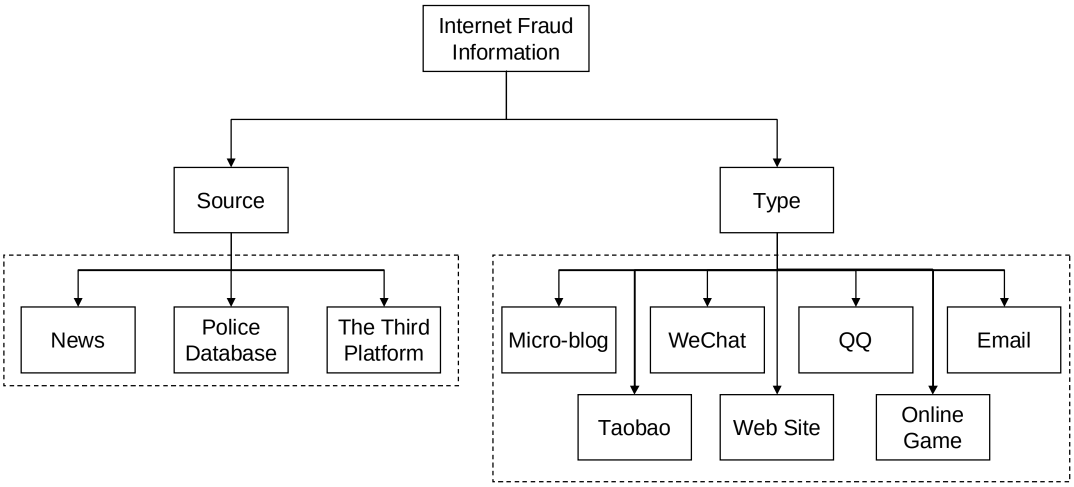

Fig. 1 Internet fraud information classification

Automatic classification model of retrieval results

The automatic classification model of the internet fraud information includes data collection, data processing and classifier. The data collection is used to solve the problem of data source, the data processing is used to process the original text and the classifier realize the function that sort the data according to the data property, as shown in Fig. 2.

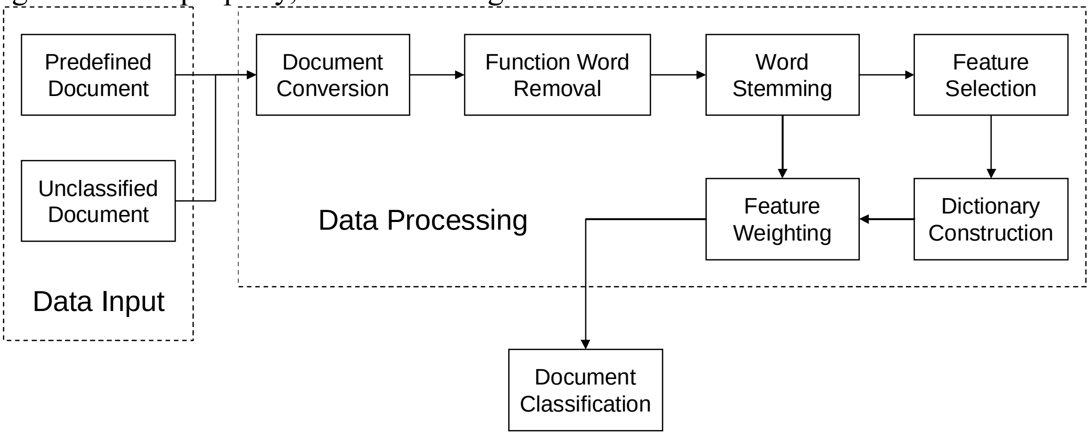

Fig. 2 Structure of scientific research information classification model

The classifier construction is the key technology of the model. This study adopts the K-NN (K-Nearest Neighbor) algorithm to classify the text, because the K-NN algorithm is efficient and simple that probably realizes the same effect as the Bayesian Classification in theory. The K-NN algorithm computes the K nearest neighbor of the search items in the training sample set, finds the class of the most of the K nearest neighbor, and then classifies the search item as the class.

The idea of classification is the data x belong to the kind of sample set that includes the most samples in the k nearest neighbor of x. The formula is as following:

$$d(x)= k_i, i=1,2,...,c, k_i = \left[\sum_{i=1}^{n} (x_i - y_i)^2 \right]^{1/2} \tag{1}$$

The rule:

$$m = \max_{i=1,2\cdots,c}(d_i(x)), x \in w_m, \tag{2}$$

The algorithm flow is as following:

Step 1. Collect and process the data.

Step 2. Design the training set and testing set.

Step 3. Set the parameter k.

Step 4. Build a queue that is sorted by the distance and store the nearest training set.

Step 5. Get randomly the k items from the training set as the initial nearest items, compute the Euclidean distance of the k items from the training set, and then store these data in the queue.

Step 6. Traverse the training set, compute the distance L of the training set from the testing set, compare the L with the Lmax, and if L>=Lmax then visit the next item else store the current item in the queue.

Step 7. Find the class that includes the most items in the queue.

Step 8. Set the different parameter k, repeat the above steps, and finally get the parameter k of the lowest error.

Performance testing

For testing the performance of the automatic classification model, this study collects 50,000 data of internet fraud, filters the distractions that will affect the classification, extracts the simple abstract of the information, establishes the semantic dictionary, builds the corresponding vector space model, and then builds the sample set for classification. The system computes the Euclidean distance between the search result set and the sample set by the K-NN algorithm and classify the search result. The testing adopts the MicroP and MicroR as the assessment index. The formula is as follows:

$$MicroP = \sum_{i=1}^{N} A_i \Big/ \left(\sum_{i=1}^{N} A_i + \sum_{i=1}^{N} B_i \right) \tag{3}$$

$$MicroR = \sum_{i=1}^{N} A_i \Big/ \left(\sum_{i=1}^{N} A_i + \sum_{i=1}^{N} C_i \right) \tag{4}$$

where MicroP is the average precision, MicroR is the average recall, N represents the total of the scientific research information classes, A_i represents the number of the correct data in the search results, B_i represents the number of the incorrect data in the search results, C_i represents the data that is correct but is not classified in the search result.

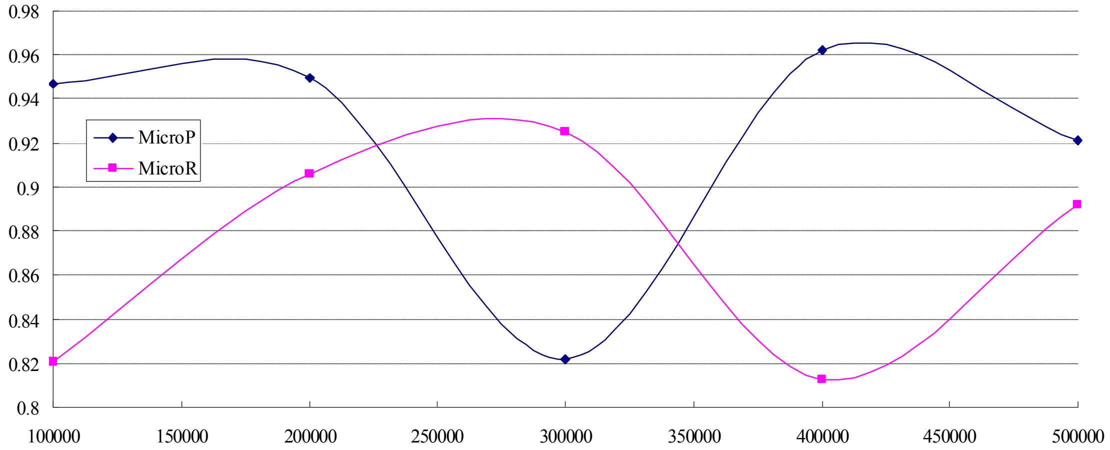

Fig. 3 Classification comparison of different samples

The search result is displayed according to the information levels in Fig. 3 and the system has higher average precision and recall. As the chart shows, the dependency of the samples amount in our system is lower than the common system because the preprocessed data structure has more semantics that increase the classification performance.

Discussion and future work

Although the Chinese internet fraud information management platform can meet the basic demands of the user, the cost of information retrieval doesn't decrease with the increasing of the acquired information quantity. For this, we build a personal retrieval platform of the internet fraud information for improving the search efficiency and decrease the cost the information retrieval of the user. In the next work, the platform will provide not only the rich information retrieval, information customization and information comparison service but also the valued report for the user through recording and analyzing the keywords search log.

Acknowledgment

This author's work is supported by the JiangXi Social Science Planning Projects during the 12th Five-Year Plan(14TQ05) and JiangXi Police College Scientific Research Project(2014QN001).

References

[1] Rocchio J. Relevance feedback in information retrieval[M]. In:The SMART retrieval system: experiments in automatic document processing. Salton G Prentice-Hall, Englewood Cliffs, 1971:313-323.

[2] RILOFF E. Automatically Constructing a Dictionary for Information Extraction Task[A].Proceeding for the Eleventh National Conference on Artificial Intelligence [C],1993.811-816.

[3] SODERLAND S. Learning Information Extraction Rules for Semi-structured and Free Text[J].Machine Learning,1999,34(13):233-272.

[4] GUO G, WANG H, BELL D. KNN model-based approach in classification[C]. On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, LNCS 2888. Berlin: Springer-Verlag, 2003: 986-996.

[5] Liu J., Huang H. The availability of search engine by classification and clustering[J]. Railway Computer Application,2006,15(3): 44-46. (in Chinese)

[6] Dai Y. The design of police decision support system based on data warehouse[J]. Micro-Computer Information, 2007, 23(6):179-180. (in Chinese)

[7] Liu Y. The preventive policy of internet fraud[J]. Information Network Security, 2008, (1):70-73. (in Chinese)

[8] Wang Z. The discussion of internet fraud[J]. ShanXi Police College Journal, 2009, 17(3):68-70. (in Chinese)

[9] Wang Z., Liu N. Classification theory research for information retrieval based on pragmatic-philosophical view [J]. Journal of Library Science,2009,31(11):1-4. (in Chinese)

[10] Yang X. The construction of internet fraud prevention system[J]. Administration and Law, 2011, (8):55-60. (in Chinese)

[11] Sun J. The research of information strategy for internet fraud[J]. Rail Police College Journal, 2013, 23(4):31-34. (in Chinese)