Application of Web Page Classification in a Domain-specific Search Engine

Chunyan Liang
North China Electric Power University
cyliang.teaching@gmail.com

***Abstract*-Automatic web page classification can be used in domain-specific search engines to help users get the specific information more conveniently and precisely on Internet. The semantic similarity and noisy data in domain-specific web pages make traditional classifier perform poorly on them. In this paper, a dictionary-based multilingual web page classification method is proposed to try to improve the classification performance. A domain-specific dictionary is constructed in the method to intensify the domain-specific knowledge in the pages. An automatic encoding detection and integration method is also introduced in the classifier to extract Chinese and English information precisely from the multilingual pages. After verified in the experiments, the method is integrated into a real domain-specific search engine where it shows good effectiveness.**

Keywords-Web page classification, Search engine, Domain-specific knowledge, Dictionary

I. INTRODUCTION

With the rapidly growth of the World Wide Web, it becomes increasing difficult to find just what we want. Domain-specific search engines could be used to help scholars and profesionals who focus on specific areas to get the specific information more conveniently and precisely on Internet. Automatic classification is an efficient techniques that can be applied in search engines to facilitate users' search [1]. However, the semantic similarity and noisy data in domain-specific web pages make traditional classifier perform poorly on them [2]. This paper proposes some methods to try to improve the classification performance.

For the domain specific web pages, the classification task becomes more challenging because they have the more similar semantic information than general pages. To solve this issue, a domain-specific dictionary is constructed to represent the domain knowledge that can be used to intensify the domain-specific knowledge in the pages [3], and a new web page classification approach based on the dictionary is proposed to attempt to classify the web pages more efficiently.

Web pages may be written in multiple languages with different encoding schemes on the Internet [4]. Without the right encoding information, the feature words of web pages will not be correctly extracted by the classifier. So, an automatic encoding detection and integration method is introduced in the web page classifier to extract Chinese and English information precisely from the mixed bilingual Chinese-English pages.

From the experiments on the real domain-specific web pages, the proposed classification method shows better performance than the traditional method. This motivates us to apply the classification method to a real domain-specific search engine.

II. DICTIONARY-BASED MULTILINGUAL WEB PAGE CLASSIFICATION METHOD

In order to recognize the domain-specific knowledge accurately, a machine-readable chemistry dictionary, named ChemDict, is first constructed to classify the chemical web pages more efficiently. ChemDict has total 172,786 Chinese terms and 173,895 English terms. ChemDict includes large amounts of phrase terms consisting of more than one word. Then, a new web page classification method is proposed based on the dictionary.

For the multilingual web pages, it is simple to recognize English characters since almost all the encoding schemes have the same code-point range for them as ASCII. It becomes hard when dealing with Chinese characters, because the commonly used encodings, such as GB2312, Big5 and UTF-8, have different code-point ranges for them. Meta HTML tags of the web pages can be used to get the encoding. For the pages without explicit charset declaration, a Character Distribution Method [4] is used in the classifier to identify the encoding based on the code point distribution statistics in the pages. Since different encodings may have different code points for the same Chinese character, it is necessary to integrate them into a uniform encoding to make the pages with different encodings, while with same meaning, have the same represented vector in the classifier. In this paper, GB2312 is chosen as the uniform encoding as it covers both the English and commonly used Chinese characters and it is easy to identify the bi-byte Chinese characters and single-byte English characters. After encoding detection and integration, all the Chinese-English pages with different encodings are integrated and can be represented accurately in the classifier.

In this paper, a modified kNN classifier is used to classify the domain-specific web pages into a hierarchical topic structure for its robust and effective performance [5]. It should be mentioned that the traditional text classification techniques could only deal with the English text.

Figure 1 illustrates the data flow of the dictionary-based multilingual web page classification method. In this method, encoding detection and integration method is first performed separately on the test and training web pages, and then automatic segmentation is done to extract

Published by Atlantis Press, Paris, France.

dictionary terms from the pages based on the chemical dictionary. After that, feature extraction is performed to produce the final vectors. Based on the training vector space, the kNN classifier finds the k nearest neighbors of the test web page, and the relevant categories of the test are finally obtained.

III. EXPERIMENTS

A. Datasets and Performance Measures.

To test the web page classification method, a labeled Chinese-English chemical web page dataset, named as ChinPage, is constructed. The pages are mainly collected based on the ChIN [6] resource links from Internet by a real-time crawler. ChinPage dataset uses a hierarchical chemical classification structure adopted by Natural Science Foundation of China. The classification structure is a three level hierarchical topic structure and has total 341 categories. Each chemical web page in ChinPage is classified into one or more categories of this structure.

ChinPage dataset is uneven or skewed multi-label dataset and has complex category structure. This means that the word/category correspondences are more complex in ChinPage dataset than other common-used datasets. ChinPage dataset is split to training/test data according to the ChIN indexed date of each page. The documents that are unlabeled or have no text besides the tags are removed from the dataset. Finally, ChinPage dataset gets 1635 documents (including 255 Chinese documents) in training set and 702 (including 40 Chinese documents) in test set.

For measuring the average performance of the classifier over multiple categories, the traditional micro-averaging F_1 [5] is adopted in the following experiments.

B. Results and Discussion.

In the experiments, the traditional classification method is first performed as a baseline run on the ChinPage dataset. Both the dictionary and the encoding detection and integration method are not applied in the modified traditional classification method. After that, the chemical dictionary is used in the classifier to extract the chemical features to observe its effect on classification performance. Then, the complete dictionary-based multilingual web page classification method that utilizes the encoding detection and integration method is tested in the classifier. The results are shown in Figure 2. The micro-averaged F_1 values are marked on the top of the columns in the figure.

From the results, we can find that the chemical dictionary can help to get the better classification performance than the traditional method when it is used to extract and strengthen the domain specific information of the web pages. The encoding detection and integration method can further improve the performance since it allows classifier to accurately identify the character encodings of multilingual pages. For the micro-averaged F_1 on all the categories, the dictionary-based multilingual web page classification method can obtain about 11% improvement over the traditional method.

IV. APPLICATION IN A REAL DOMAIN-SPECIFIC SEARCH ENGINE

To observe the actual effects on the domain-specific search engines, we integrats the proposed dictionary-based multilingual web page classification method into a real chemical search engine, which is called ChemEngine. ChemEngine is a prototype of chemistry-focused search engine, developed to help chemists find chemical information more conveniently and precisely on Internet. ChemEngine indexes millions of multilingual chemical web pages.

All of the web pages in ChemPage dataset are used to train the classifier, and then the web pages indexed in ChemEngine are classified and assigned into the three level hierarchical classification structure. The relation between the pages and the categories is preliminarily stored in the background database.

A subject tree representing the hierarchical category structure is offered in addition to a normal result list when a user submits a query, for example “benzene”, to ChemEngine as Figure 3 shows. Once the user clicks a node of the tree, the corresponding web pages of the category will be selected from the origin result list.

We can observe the acute reduction of the result page number when a category is chosen to refine the search. It helps user to get the needed information more easily and quickly.

V. SUMMARIES

Automatic classification of the domain-specific web pages into a hierarchical topic structure is a difficult issue. The proposed dictionary-based multilingual web page classification method can effectively improve the classification performance as shown in the paper, and can provide quite accurate classification results to satisfy users’ need when applied in ChemEngine, a chemistry-focused search engine.

ACKNOWLEDGMENT

This work is sponsored by the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation of China under grant No.71071054.

REFERENCES

[1] A. Selamat, S. Omatu, Web Page Feature Selection and Classification Using Neural Networks, Information Sciences [J], 2004, vol.158, PP:69-88

[2] W. Huang, L. Xu, J. Duan, Y. Lu, Chinese Web-page Classification Study, IEEE International Conference on Control and Automation [J], 2007, vol.1-7, PP:2141-2146

[3] C. Quan, F. Ren and T. He, Sentimental Classification Based on Kernel Methods and Domain Semantic Orientation Dictionaries, International Journal of Innovative Computing, Information and Control [J], 2010, vol.6, no.6, PP:2681-2690

[4] S. Li and K. Momoi, A Composite Approach to Language/encoding Detection, Nineteenth International Unicode Conference [C], San Jose, California, 2001

[5] Y. Yang, An Evaluation of Statistical Approaches to Text Categorization, Information Retrieval [J], 1999, vol.1, PP: 69-90

[6] ChIN: The Chemical Information Network, Institute of Process Engineering, Chinese Academy of Sciences, http://chin.csdl.ac.cn/, 2011.

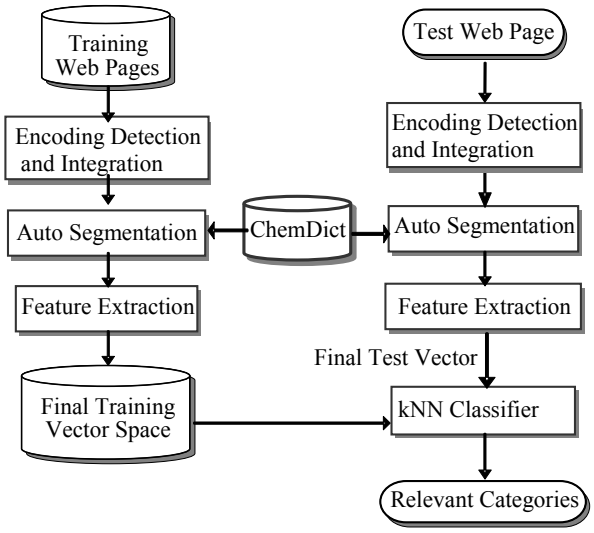


Figure 1. Dictionary-based multilingual web page classification method

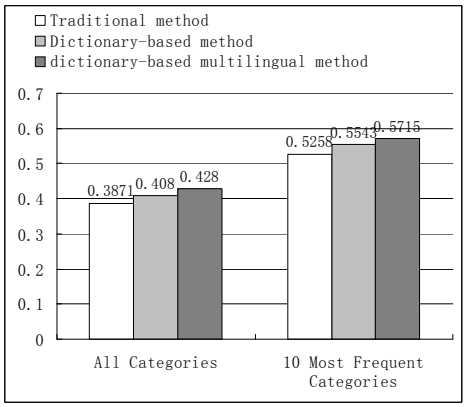
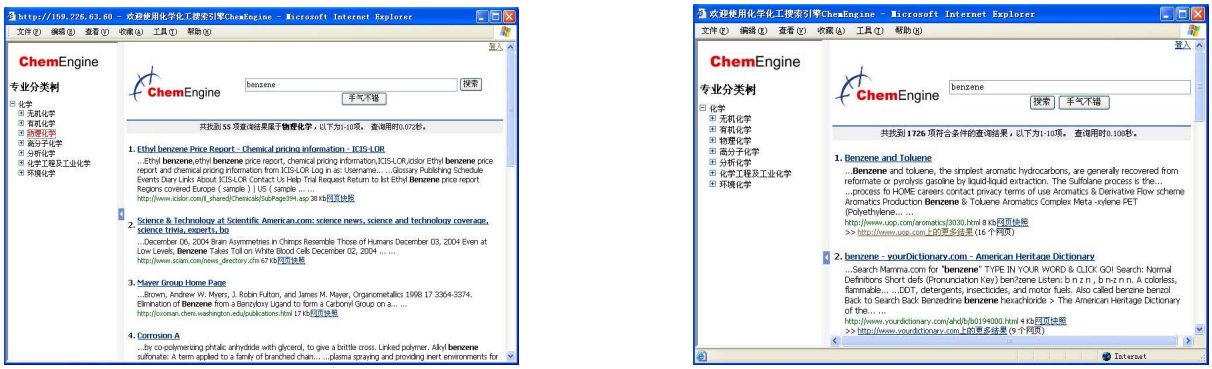


Figure 2. Classification performances of the different classification methods on ChinPage dataset



(a) (b)

Figure 3. The result interface in ChemEngine of query “benzene” ((a) all search results, (b) classification results)