An Improved Approach to Term Weighting in Hierarchical Web Page Classification

Jinbo Tan[1*]

[1] Educational Technology Department, Shandong Normal University, Jinan, Shandong Province, China.

* Corresponding author. Tel.:13589079152; email: yttjb@163.com
Manuscript submitted January 10, 2014; accepted August 8, 2014.

Abstract: Currently, in web page classification, Absolute Weighting Method is a common method to weight HTML main structure features. The disadvantage of the method is that weighting coefficient is a fixed value, which has different effects on the long and short text. So the influence of structure features on local text will be weakened with the length of local text increasing. To solve the problem, we propose an improved weighting method, namely Relative Weighting Method. In the experiment of web page hierarchical classification, we compare the two methods' classification performance on a single label and several labels combination. The results show that Relative Weighting Method can effectively improve the classification accuracy, which is better than the Absolute Weighting Method.

Key words: Hierarchical web page classification, relative weighting method, term weighting algorithm, web structure.

1. Introduction

At present, a mass of structured information are involved in web page. For example, title, hyperlinks, HTML marks and metadata information have been applied to web page classification technology [1]-[7]. Most Web pages have a title element enclosed by the <title> and </title> tags. The title of a page usually gives a good summary of the content of the page, where there are basically key words except a few function words [8]. Metadata used as index provides generalization of website content for searching engine, which reflects the key information of the web page in author's view, but can't display in the browser. Different from the aforementioned two kinds of tag domains, hyperlink anchor text has its own characteristics. Hyperlink anchor text is the descriptive text corresponding to the link which one web page contains, and which can redirect to another one. For example, there is a link like ' 小学语文教学' , it is clear that '小学语文数学' is the hyperlink anchor text of the address 'http://www.vastman.com'. In the classification based on hyperlink anchor text, there is a basic assumption: hyperlink anchor text is mainly used to describe the document it point to and not the document where it is in [9]. Some has been done to improve the accuracy of web page classification using hyperlinks [10]-[14]. As for HTML documents, we usually combine several tag domains rather than only one to improve classification precision rate and recall rate in most cases, because one tag domain alone contains less or incomplete information.

Absolute Weighting Method is the common method to weight the context features of web pages, which can improve web page classification effects to a certain extent. But because weighting coefficient is a fixed

value, which has different effect on the long and short text, the influence of context features to local text will be weakened with its length increasing. Besides, most researches are involved with flat classification instead of hierarchical classification [15]-[19]. Hierarchical classification researches only study English web pages as well [20].

For Chinese web page, when categories are organized in tree structures, can structure features of web page improve hierarchical web page classification effects? What are the disadvantages of the common Absolute Weighting Method to hierarchical web page classification? How to improve the method? These are the main subjects to be discussed in this paper. The following part of the paper will expatiate on the representation of web page. Then, the Relative Weighting Method will be presented in the 3rd part and the method is confirmed by the experiments on the Internet elementary education corpora in the 4th part. At last, the summary of the paper will be given in the 5th part.

2. Web Page Feature Representation

Web pages are essential hypertext. Besides text and multimedia components, Web pages also include context features such as hyperlinks, HTML tags and metadata [21]. Various context features have been used in Web classification. According to the various use of context features and different ways these context features derived, Web page classification can be categorized into text only, hypertext and relational learning approaches. Nevertheless, most of the experiments were conducted on flat category spaces. When the categories are organized in a tree-like structure, we study the impact of the context features on hierarchical Web page classification. We can use content features alone as the baseline features and try out different combinations of page-derivable context features in top-down level-based classification methods. The Web page feature combinations are given below.

2.1. Local Text

In this combination, each Web page is represented by using a set of words extracted from its text component only [8]. The text component of a page is obtained by removing HTML tags from the page. Suppose there are X unique words in the collection. A Web page p is then represented by a binary feature vector derived from its corresponding plain text document, $p = \{x_1, x_2, ..., x_X\}$, where x_i equals 1 if the corresponding word occurs in the text document, or otherwise 0.

2.2. Text + Title

In this representation, words extracted from the title element of page are used as features in addition to the features from the text component. Let t_j be the binary feature associated with a unique word observed in the titles of all the training Web pages. Suppose there are T such unique words. A Web page p is represented by a binary feature vector from its plain text component and title element, $p = \{x_1, x_2, ..., x_X, t_1, t_2, ..., t_T\}$. Note that a word appearing in the title element also appears in the text document extracted from the Web page. Therefore, two features ids will be assigned to each word.

2.3. Text+ Metadata

In this representation, words extracted from the metadata element of page are used as features in addition to the features from the text component. Let m_k be the features associated with metadata, supposing there are M such words. A Web page p is represented by both the local words and metadata words, $p = \{x_1, x_2, ..., x_X, m_1, m_2, ..., m_M\}$.

2.4. Text + Anchor Words

Anchor words associated with an in-link to a Web page carry some basic descriptions about the page or certain parts of the page. We therefore consider the in-link anchor words as collection-derivable context features in Web page representation. Let a_k be the binary feature associated with a unique word observed in anchor words which are associated with the links pointing to any of the Web pages in a training set. Suppose there are A such words. A Web page p is represented by both the local words and anchor words, $p = \{x_1, x_2, ..., x_X, a_1, a_2, ..., a_A\}$. Similarly, a word which appears in both the local text component and the in-link anchor words is assigned two different feature ids.

2.5. Text + Title + Anchor Words

To study the effect of different context features, a Web page is represented by all features, i.e., local text, title words and in-link anchor words, i.e., $p = \{x_1, x_2, ..., x_X, t_1, t_2, ..., t_T, a_1, a_2, ..., a_A\}$.

3. Improved Weighting Method Based on Web Page Structure

Document is usually denoted by vector space model, and term weighting formula is denoted by TF.IDF as follows:

$$W_{ij} = \frac{tf_{ij} \log(\frac{N}{n_j} + 0.01)}{\sqrt{\sum_{j=1}^{n} (tf_{ij} \log (\frac{N}{n_j} + 0.01))^2}} \quad (1)$$

Different from the plain text document, HTML document contains abundant identifiers which make the context structure more distinct. According to the significant subject information included in title, hyperlink anchor words and Metadata, tf_{ij} in the formula above can be changed as follows:

$$tf_{ij} = \sum_{e_k} (w(e_k) \cdot tf(t_i, e_k, d_j)) \quad (2)$$

e_k: web page elements such as body text, title, anchor text; $w(e_k)$: e_k weighting; tf (t_i, e_k, d_j): the frequency of t_i in element(e_k) of HTML document(d_j).

Generally, a certain absolute value is given to α to add the homologous element weighting. Take local text and title method for example, we can scale-up the title words of all test web pages by α times, then weight them together with the local text words. But the problem is that α is a fixed value which has different effects on long and short text, and the influence of the title on local text will be weakened as its length increasing. For example, there are 1000 words in the local text of HTML document A and 100 in B, and both A and B have 4 terms in their titles. Then, when α is 5, title terms number is enlarged to 20, these 20 words have less influence on A than that on B.

To solve this problem, the paper proposes an improved weighting method called Relative Weighting Method whose formula is as follows:

$$w(e) = \begin{cases} \frac{sum(d_j)}{sum(e)} \cdot \frac{\alpha}{1-\alpha} & e \text{ is anchor or title or metadata terms, } \alpha \text{ is an adjustable parameters between 0-1;} \\ & sum(d_j): \text{ the number of terms in dj, } sum(e): \text{ the number of terms in e} \\ 1 & e \text{ equals 0} \end{cases} \quad (3)$$

In Relative Weighting Method, the length of HTML document local text and its context terms is taken into consideration. The amount of context terms is enlarged in proportion according to the length of HTML document and the influences of these terms on local text are less affected by the document length.

The details are as follows: first, the words of the test HTML document are segmented, and the statistics of local text and context terms are taken separately, then determine the total amount of words according to α. Still take local text and title as an example, assume that there are 1000 words in the local text of HTML document A and 100 in B, and both A and B have 4 words in their titles. Then, when α is 0.1, title word number of A is enlarged to 112 and B enlarged to 12, so the influences of the two cases are identical.

4. Experiment

This section gives the details of the proposed research on automatic web page classification of educational resources. First, experimental environment is introduced. Then, it shows the experimental results.

4.1. Environment and Data

To perform these feature selection methods in the hierarchical web classification, we take elementary education subjects as the study object, and build a hierarchical system for 12 parent-categories and 18 sub-categories in Fig. 1. We build a dataset of 7412 training documents and 1100 test documents manually, as shown in Table 1.

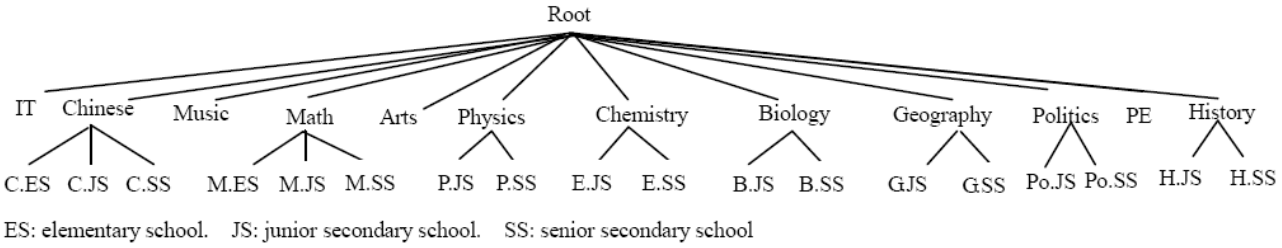

Fig. 1. Hierarchical category of elementary education subjects.

Table 1. Training Documents and Test Documents

Category	Training	Test	Category	Training	Test
M.ES	300	50	G.SS	265	50
M.JS	300	50	B.JS	438	50
M.SS	300	50	B.SS	357	50
C.ES	300	50	H.JS	281	50
C.JS	295	50	H.SS	302	50
C.SS	300	50	Po.JS	299	50
P.JS	228	50	Po.SS	317	50
P.SS	324	50	Music	300	50
E.JS	290	50	Arts	499	50
E.SS	394	50	PE	500	50
G.JS	327	50	IT	496	50

We adopt improved mutual information algorithm as feature selection method [22], use TF.IDF algorithm as weight computation [23], and use the Rocchio algorithm as the classification method [24]. When training, the positive training documents for each parent-category are composed by its sub-categories' documents, and the negative training documents are selected from the documents belonging to the sub-tree at the parent level. The negative training documents for each sub-category are composed by other sub-categories'

documents under the same parent-category. Finally, we select 3000 features for each category. When testing, we use top-down level-based classification method, that is, web page is first classified into a parent-classifier, and then judged by each sub-classifier under the parent-category whether it belongs to its sub-category. The performance measure in classification uses macro/micro-averaged precision, recall and F1.

4.2. Experimental Results

We study the results from three aspects 1) what are the impacts of different Web page representations (context features) in Absolute Weighting Method and Relative Weighting Method on hierarchical classification. 2) which method performs better on elementary education subject dataset. 3) Compare the effects of Relative Weighting Method in flat classification and hierarchical classification.

The macro-averaged precision and recall measures of Absolute Weighting Method are reported in Fig. 2 and Relative Weighting Method in Fig. 3.

From the Fig. 2 we can see that metadata has little effect on the classification results. The reason is that most of the HTML documents have no metadata, while only 224 pages have metadata in 1100 test web pages. Part of the metadata contains advertisements which can't explain the web theme. So we don't consider metadata when performing the Relative Weighting Method. Compared with using the local text only, inclusion of feature sets having either title or anchor leads to improvement in both macro-averaged precision and recall. When α equals 36, the two measures both reach the maximum value, however, the macro-averaged precision is better than recall.

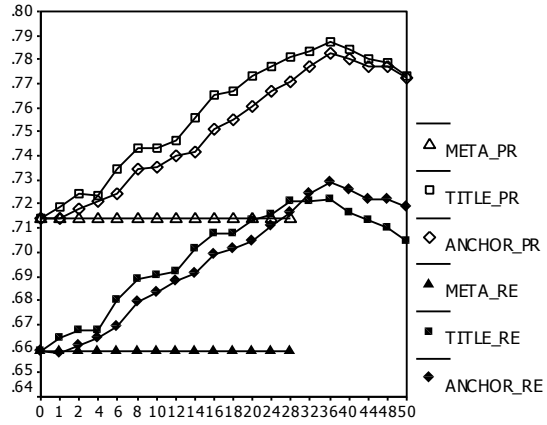

Fig. 2. Macro-averaged measures on α being absolute value.

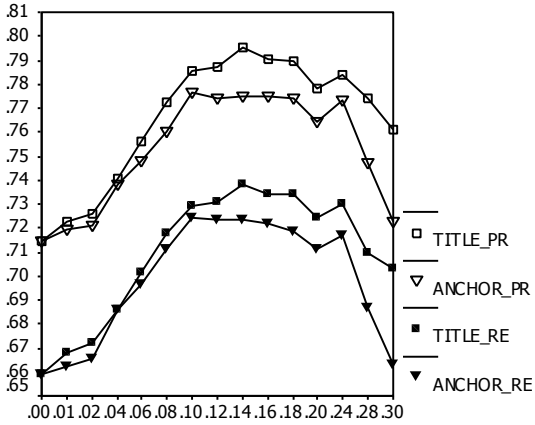

Fig. 3. Macro-averaged measures on α being relative value.

In the Relative Weighting Method, the trend of macro-averaged precision and recall is similar to that in Absolute Weighting Method. For the title, when α equals 0.14, the highest macro-averaged precision and recall are achieved. The same situation appears when α equals 0.10 for the anchor. From these two sets of performance comparison, it is clear that the context features including anchor and title had positive effects on Web page classification especially on precision.

To compare the absolute and Relative Weighting Methods, the performance of using text + title, text + anchor, text + title + anchor of the two methods are plotted in Fig. 4. Note that in the experiment each α equals the maximum value in the Fig. 2 and Fig. 3. When both title (α equals 0.13) and anchor (α equals 0.11) considered, better results on macro-averaged recall, micro-averaged precision, recall and F1 are observed with the relative weighting method. Overall, according to the measure results, the Relative Weighting

Method is better than the Absolute Weighting Method, and the effect of text + title + anchor is better than the other two combination methods.

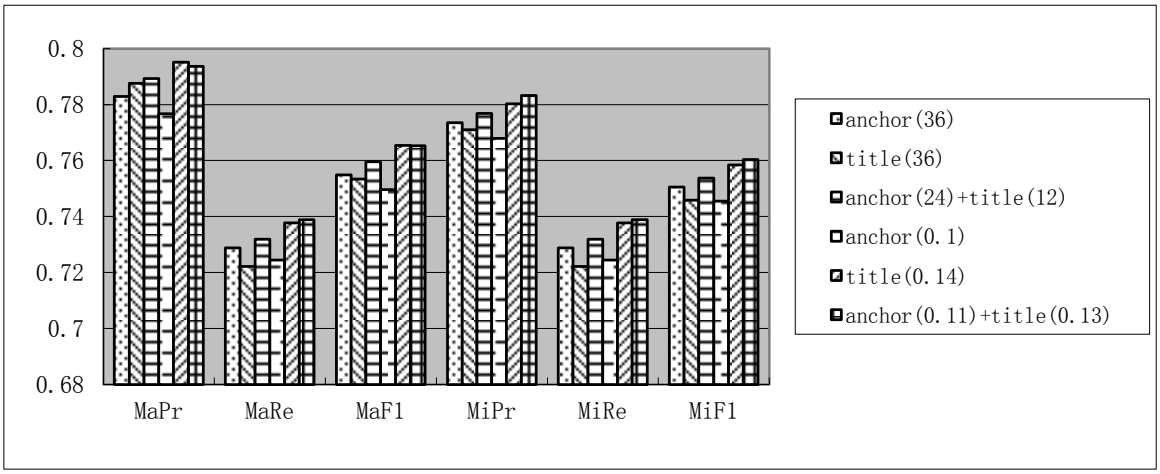

Fig. 4. Overall performance of absolute and relative weighting method using features anchor and title.

To verify the Relative Weighting Method, the comparison of it to flat classification and hierarchical classification has been performed; the results are presented in Table 2. It should be noticed that α is a relative value which is the maximum value of the above tests. The Relative Weighting Method performs better in hierarchical classification than in flat classification. Especially, the representation which contains the title or anchor text can improve the macro-averaged precision, recall and F1 values.

Table 2. Macro-Averaged Values of Flat Classification and Hierarchical Classification

Classification method	Flat Classification			Hierarchical classification		
Web features (α)	Macro_Pr	Macro_Re	Macro_F1	Macro_Pr	Macro_Re	Macro_F1
Local text	0.691	0.636	0.663	0.714	0.659	0.685
Local text + title (0.14)	0.746	0.692	0.718	0.795	0.738	0.765
Local text + anchor text (0.10)	0.723	0.678	0.700	0.777	0.724	0.750
Local text + anchor text(0.11) +title(0.13)	0.731	0.682	0.705	0.794	0.739	0.765

5. Conclusions

In the paper, we studies two weighting methods: Absolute Weighting Method and Relative Weighting Method. Our experiments showed that the Relative Weighting Method we proposed is better than Absolute Weighting Method usually used, because the Relative Weighting Method can't be influenced by the documents' length. The impact of context features on the Web page classification was studied with a hierarchical category tree derived from elementary education subjects' dataset. We experimented the combinations, text with title, text with metadata, text with anchor words and text with both title and anchor words on the hierarchical classification. Our experiments showed that the use of context features can lead to striking increase in precision of web classification.

Acknowledgement

This work is supported by the youth project of Education Ministry for national education sciences planning "Research on learners' behavior characteristics and strategies construction of searching

information online" (No. ECA130376).

References

[1] Robert, C., Bamshad, M., & Jaideep, S. (1997). Web mining: Information and pattern discovery on the world wide web. *IEEE Transactions on Applications and Industry* (pp. 558-567).

[2] Chakrabarti, S., Dom, B. E., & Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. *Sigmod Record (ACM Special Interest Group on Management of Data)*, *27(2)*, 307–318.

[3] Attardi, G., Gull, A., & Sebastiani, F. (1999). Automatic web page categorization by link and context analysis. *Proceedings of THAI-99, European Symposium on Telematics, Hypermedia and Artificial Intelligence* (pp. 1-16).

[4] Rayid, G., Seán, S., & Yiming, Y. (2001). Hypertext categorization using hyperlink patterns and meta data. *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 178-185).

[5] William, W. C. (2002). Improving a page classifier with anchor extraction and link analysis. *In Advances in Neural Information Processing Systems* (pp. 1481-1488).

[6] Glover, E., Tsioutsiouliklis, K., Lawrence, S., Pennock, D., & Flake G. (2002). Using web structure for classifying and describing web pages. *Proceedings of the World Wide Web* (pp. 562-569). Hawaii, USA.

[7] Oh, H. J., Myaeng, S. H., & Lee, M. H. (2000). A practical hypertext categorization method using links and incrementally available class information. *Proceedings of Research and Development in Information Retrieval – SIGIR* (pp. 264-271).

[8] Sun, A., Lim, E. P., & Ng, W. K. (2002). Web classification using support vector machine. *Proceedings of the 4th International Workshop on Web Information and Data Management* (pp. 96-99).

[9] Blanco, L., Crescenzi, V., & Merialdo, P. (2008). Structure and semantics of data-intensive web pages: An experimental study on their relationships. *Journal of Universal Computer Science*, *14(11)*, 1877-1892.

[10] Choi, B., & Yao, Z. (2005). Web page classification. *Studies in Fuzziness and Soft Computing*, *180*, 221-274.

[11] Inma, H., Carlos, R. R., David, R., & Rafael, C. (2011). A tool for link -based web page classification, advances in artificial intelligence. *Lecture Notes in Computer Science*, *70(13)*, 443-452.

[12] Sun, A. (2004). Ontology-based web classification. Unpublished undergraduate dissertation. Nanyang Technological University, Singapore.

[13] Michael, C., & Hsinchun, C. (2008). A machine learning approach to web page filtering using content and structure analysis. *Decision Support Systems*, *44(2)*, 482-494.

[14] Sukakanya, U., & Porkaew, K. (2011). Modeling a generic web classification system using design patterns. *Journal of Computers*, *6(10)*, 2212-2220.

[15] Furnkranz, J. (1999). Exploiting structural information for text classification on the WWW. *Proceedings of the 3rd Symposium on Intelligent Data Analysis* (pp. 487-498).

[16] Daniele, R. (2002). Feature selection for web page classification. *Proceedings of the IEEE – PIEEE* (pp. 102-116).

[17] Yang, Y., Slattery, S., & Ghani, R. (2002). A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems*, *18(2-3)*, 219-241.

[18] Sun, A., Lim, E., & Wee, K. (2002). Web classification using support vector machine. *Proceedings of the 4th International Workshop on Web Information and Data Management* (pp. 96-99).

[19] Vidal, M., Silva, A. S., Moura, E. S., & Cavalcanti, J. M. B. (2008). Structure-based crawling in the hidden web. *Journal of Universal Computer Science*, *14(11)*, 1857-1876.

[20] Miao, Y. Q., & Kamel, M. (2011). Pairwise optimized rocchio algorithm for text categorization. *Pattern Recognition Letters*, *32(2)*, 375-382.

[21] Aggarwal, P., Vig, R., & Sardana, H. K. (2013). Patient-wise versus nodule-wise classification of annotated pulmonary nodules using pathologically confirmed cases. *Journal of Computers*, *8(9)*, 2245-2255.

[22] Tan, J. B. (2006). *Study of Automatic Classification Technology for Online-Based Elementary Education Resource*. Nanjing: Nanjing Normal University.

[23] Ding, L., Yu, F., Peng, S., *et al.* (2013). A classification algorithm for network traffic based on improved support vector machine. *Journal of Computers*, *8(4)*, 1090-1096.

[24] Hinrich, S., David, H., & Jan, P. (1995). A comparison of classifiers and document representations for the routing problem. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 229-237).

Jinbo Tan was born in 1978. She received the Ph.D. degree in 2006 from Nanjing Normal University. Now she is an associate professor in School of Communication of Shandong Normal University. She has published 14 papers indexed by CSSCI and 3 papers indexed by EI. Her research interests include data mining and information retrieval.