2016 International Conference on Computational Modeling, Simulation and Applied Mathematics (CMSAM 2016)
ISBN: 978-1-60595-385-4

A Method of Web Page Classification Based on Feature Dimension Reduction

Xun-yi REN* and Dan ZHANG

Nanjing University of Posts and Telecommunications, Nanjing, China

*Corresponding author

Keywords: Web page classification, Bloom filter, Feature weight, Naive bayes.

Abstract. Text classification technology to quickly retrieve information pages, information filtering and data mining provides an important foundation. Due to the diversity and complexity of web page format, the problem of web page classification is more difficult to deal with than text classification. This paper on web page classification method research, through the improvement of bloom filter, after pretreatment of the text content of pages, the improved bloom filter used in filtering feature set, greatly reducing the feature dimension, then according to the characteristics of web page, the feature weight algorithm has been improved. Finally, using the naive Bayesian classifier can verify the effectiveness of the algorithm.

Introduction

Web page classification refers to web pages according to the classification standard or system of automatic classification marking. Research on Web page classification technology is gradually becoming after the research of text classification machine learning research hotspot in the field, one of the main applications of web page classification technology is also a text classification, text classification technology for web information fast retrieval, the information filtering and data mining provide the important basis [1]. Due to the diversity of web page format, the content is complex, in addition to pure text, there are many other contents on the classification, which leads to the problem of web page classification is more difficult to deal with than text classification. The process of web page classification is actually the process of collection, pretreatment, text representation, training and final classification of web pages [2]. The process can be expressed in Figure 1.

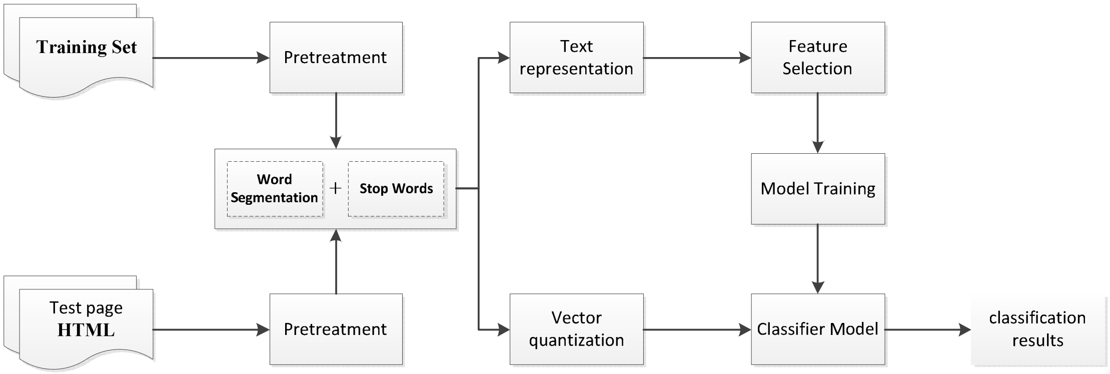


Figure 1. Web page classification flow chart.

In this paper, on the basic of the bloom filter model is improved, and its application to filtering features, which greatly reduces the feature dimension, so as to shorten the time of word frequency statistics, then use the improved feature weighting algorithm of feature weight calculation, feature vector is determined according to weight and select the naive bayesian classifier model training, finally through the experiment on the classification results are evaluated.

Improved Bloom Filter Algorithm

Bloom filter is the use of space superiority hash tables designed data structure by allowing a small number of errors to save a lot of storage space. The core of bloom filter algorithm is using multiple different hash functions to solve the conflict [3]. This paper bloom filter improvements, the goal is to reduce the error rate, but also can't let the loss of space advantage. In order to discuss the performance of bloom filter, we first need to introduce a few variables as shown in table 1:

Table 1. Bloom filter variable table.

Symbol	Meaning
P_{error}	False positive
V	Representation bit vector
N	Characterization of the size of a bit vector (bit)
k	Number of Hash functions
n	Need to represent the size of the data set
h_i	i -th Hash Functions
p_i	After n data into bit vector V , the probability of the i -th bit is 1.
T	Average decision time of search
t	The time required for a Hash value comparison

Improved Bloom Filter Algorithm Principle

False positive rate and median group size and the number of hash functions are related to the general application, the range of characterization bit vector N range than the source data set represents the number of n to be much larger, from the data x, after hash functions h, map place hash vector process must exist more conflicts, more data can be mapped to the same bit vector address, which is a hash representation of the characteristics of the basic bloom filter algorithm can reduce conflict by choosing the hash function, unless vector selection of size greater than or equal to the spatial data sets and can find a complete hash function such that its value is not repeated in the vector space addresses [4].

Improved bloom filter algorithm core is between the overall amount of space occupied and the local computation tradeoff can maintain global consistency and characterization bit vector data source collection number [5]. The basic idea of the improved bloom filter algorithm is as follows:

Use two bit vectors to represent the set of hash, one of which is the N bit size V_1, the other is the N/r size of the V_2;

Division of k hash functions into two parts k_1 and k_2. Map for V_1 is k_1, and k_2 is mapped for V_2;

When inserting an element, k_1 maps the element to the V_1 and sets the V_1 corresponding to 1. k_2 maps the data elements to the V_2 and sets 1 in the corresponding position at V_2;

If a position has been set many times, then only the first time will play a role, there will be no effect on the back several times;

Determination of elements, were examined after mapping of k_1 and k_2, V_1 and V_2 related position is 1 or not, if the total is 1, is that the element belongs to the set.

Key Formula of Improved Filter Bloom

Formulaic representation Improved bloom filter algorithm as follows:

$$V_{(i,j)} = \begin{cases} 1, \ if \ \exists x \in A, \ i=1, \ \exists r = 1,2,\cdots k/2, \ h_r(x) = j \\ 1, \ if \ \exists x \in A, \ i=2, \ \exists r = (k/2)+1,\cdots,k, \ h_r(x) = j \\ 0, \ others \end{cases} . \quad (1)$$

Among them, $V(i,j)$ represents the first j bit in the i bit vector. Each bloom filter improved bloom in the filter as part of the bloom filter in the whole operation. So we can get:

$$P_1 = 1-\left(1-\frac{1}{N}\right)^{\left(n\times\frac{k}{2}\right)} . \tag{2}$$

$$P_2 = 1-\left(1-\frac{r}{N}\right)^{\left(n\times\frac{k}{2}\right)} \tag{3}$$

Next, according to the multiplication principle, it is easy to deduce that the improved bloom filter error rate determination:

$$P_{error}^{'} = P_1^{k/2}\times P_2^{k/2} = \left(1-\left(1-\frac{1}{N}\right)^{\left(n\times\frac{k}{2}\right)}\right)^{k/2}\times\left(1-\left(1-\frac{r}{N}\right)^{\left(n\times\frac{k}{2}\right)}\right)^{k/2} \tag{4}$$

Mean decision time:

$$T^{'} = \frac{1-p_1^{k/2}}{1-p_1}\times t+\frac{1-p_2^{k/2}}{1-p_2}\times t \tag{5}$$

By the above formula, we can see that the improved bloom filter theory reduces the basic bloom filter error discrimination rate, under the dynamic data set growth and improved bloom filter to the local node increase the space and time overhead and does not increase the overall space overhead in the cost of maintaining a miscarriage of justice rate. Page classification applications, under normal circumstances, we encountered was stable and slow growth of data, improved bloom filter can solve the above problem.

Improved Feature Weight Description and Calculation

Feature Weight Algorithm

The calculation of feature weights is the key factor of vector similarity calculation, directly affect the accuracy of the calculation results [6]. TF-IDF is a kind of method to calculate the normalized frequency, *tf* represents the word frequency (term frequency), which refers to a certain feature in a classification of the frequency in the text of different categories, which is a big difference in the frequency of features [7]. IDF (inverse document frequency inverse document frequency), is all the training text number already contains the removal characteristics of the number of files, in order to make the divisor impossible for 0.

The *tf-idf* algorithm to reflect the ability to distinguish between a feature item to a certain extent, but only to consider the number of occurrences of feature item appears in the text as well as the feature item in the training set frequency, ignoring the internal factors and the semantic class feature items of distribution information, does not accurately reflect the importance of the feature item in the text. Therefore, this article from the feature in the text in the position to infer semantic feature, and modified *tf-idf* algorithm, the feature weight calculation method is more reasonable, in order to improve the accuracy of text classification effect.

Improved TF-IDF Formula

Different positions in the word is not the same as the embodiment of “the theme of the role”, so the contribution degree of text classification is not the same. A web page title and subtitle first concise and comprehensive performance of the web page theme, “the top often highlighted topic, elaborating the theme", summary statements appear in web content page [8]. The rules show that the features in different positions, and its roles are not the same, although some of the features of *tf* is not high, it can well reflect the content of the text. Therefore, according to the characteristics of different positions were weighted. According to the statistical data and experience of people, set the weight coefficient of p features, as shown in table 2.

Table 2. Feature Weight Coefficient Table.

The feature position in the text	weight coefficient(p)
Title	1.0
Headers	0.8
Footers	0.6
Other position	0.3

Let l_{t_i} be the number of features t_i appears in the corresponding position, p_{t_i} position weight coefficient feature weight calculation method of t_i, introduces the position weight coefficient formula is as follows:

$$w_{ik} = w_{ik} * \frac{\sum l_{t_i} * p_{t_i}}{l_{t_i}} \quad . \tag{6}$$

Experimental Results and Analysis

Experimental Data

Corpus of this article is written in multiple crawlers to collect 12000 url, use the environment as shown in table 3:

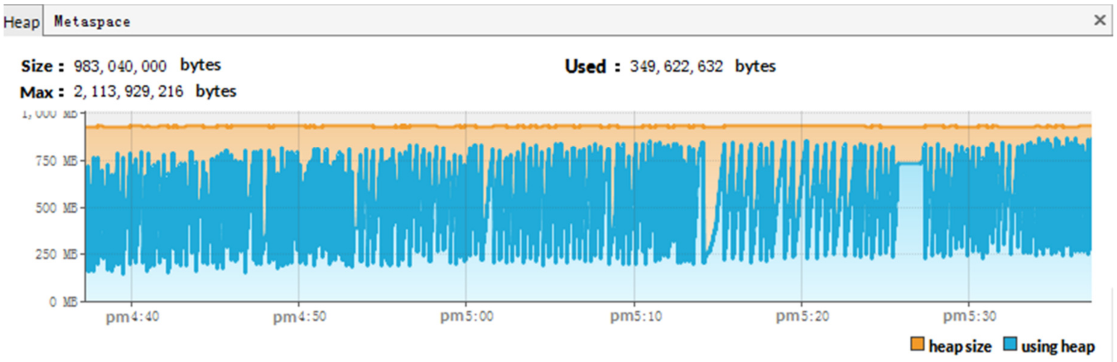
Table 3. Write the environment of reptiles.

System	Windows 7	CentOS 6.5
Environment	JDK 1.7	Python 2.6.6
IDE	Eclipse Release 4.2.0	PyCharm 4.5.1
DB	MySQL Ver 14.14	Distrib 5.1.73

Crawler program using java and python prepared in the experiment, first is to use the java to grab the page, but when to visit with protective measures, it will be rejected, but with their browser to access it, is accessible, so we use python to imitate oneself is a browser, then go to these pages for a visit, the concrete realization of the code in java and python will not go into details. The collected data and then using python for data analysis, and text content on the web pretreatment, namely to stop word segmentation and the others.

Experimental Process and Result Analysis

The text content of the web pages after pretreatment with semantic characteristics of the features is extracted. Due to the collected web pages are a large number of feature dimension that is a geometric progression growth. This will lead to an increase in the naive bayesian classifier on the induction and learning time, occupation of space will increase, and for complex degree, its growth is a geometric progression growth. Here you can completely reflect the bloom filter effect of dimension reduction, that we use improved bloom filter to the high-dimension feature sets are processed, has no intention of word meaning and meaning light words of text classification useless word removal, improved bloom filter operation after 3 hours of memory usage as shown in Figure 2, the complex space and time complexity are $O(n)$ and $O(1)$.



Size : 983, 040, 000 bytes
Max : 2, 113, 929, 216 bytes
Used : 349, 622, 632 bytes

Figure 2. Memory usage in bloom filter runtime.

After using the improved bloom filter feature set effect is remarkable, the characteristics of each category are semantic, is all contribution to classification of web page, then use improved feature weighting algorithm of all the features of the calculation, calculate each text feature item weight, each text representation for vector $(t_1, w_1; \cdots, t_i, w_i; \cdots, t_m, w_m)$, t_i as feature terms, w_i for the corresponding weights, this text from a web page with a point in a vector space expressed. Finally, using the naive bias classifier for text categorization, the experimental data show that the average classification accuracy of 89.1% or more.

Acknowledgement

Supported by National Natural Science Foundation of China (61073188).

References

[1] Zhang Xiang, Research and implementation of a web page classification system. Beijing University of Posts and Telecommunications, 2013.

[2] Battit i,R.U singmutual information for selecting feature sinsupervise dneuralnet learning. IEEE Transactionson Neural Networks, 2010, 5(4): 537-550

[3] Pei Songwen, Wu Baifeng. Research on multi class text classification algorithms based on dynamic adaptive feature weights. Computer application research, 2011, 28 (11): 4092-4096.

[4] Liu Wei, Guo Yuanbo, Huang Peng. The multi-dimensional bloom filter pattern matching engine based on. Computer Application, 2011, 31 (1): 107-109.

[5] Wang Zheng. The implementation and application of bloom filter algorithm based on the technology of removing duplicated web pages. Beijing: Beijing Jiaotong University, 2010.

[6] Naveenkmar N, Batri K. An empirical study on term weights for text categorization. International Journal of Advanced Information Science and Technology, 2012, 11:43-46.

[7] He Min, et al. Research on the application of MapReduce based on the average multinomial naive Bayes text classification. 2016, 33 (1): 115-117.

[8] Xu Linbin. Research and application of distributed web page automatic classification algorithm based on bayesian. Beijing University of Posts and Telecommunications, 2015.