International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010
1793-8201

A combination approach for Web Page Classification using Page Rank and Feature Selection Technique

Sini Shibu[1], Aishwarya Vishwakarma[2] and Niket Bhargava[3]

***Abstract*—With the rapid increase in the number of Internet users, there is a constantly increasing requirement for automatic classification techniques with greater classification accuracy. Existing algorithms are based on the text content of the web pages for classification. However, the web pages classified using text content may not yield desired results if the keywords are irrelevantly specified. This provides scope for implementing feature reduction techniques with page ranking on web pages so as to get desired results within less time and greater accuracy. Only then information on the web can be used effectively by the users. In this paper, we propose a unique method for classification of web pages by applying feature selection technique along with page rank. The aim is to improve the classification and search time of web contents so as to provide accurate and speedy results to the users.**

***Index Terms*—Web Page Classification, Feature Selection, Feature Reduction, Page Rank.**

I. Introduction

With the rapid growth of the World Wide Web (WWW), there is an ever increasing need to provide automated assistance to Web navigators for Web page classification and categorization. This assistance is helpful in organizing the vast amount of information returned by keyword-based search engines, or in constructing hierarchical collections of web documents like Yahoo directory and the MSN directory.

A. Web Page Classification

This involves classifying the web pages based on various parameters such as text, image, structure of the document etc. The classification mechanism can be classified into the following broad categories:

1) Manual classification by domain specific experts.
2) Clustering approaches.
3) META tags (which serve the purpose of document indexing).
4) A combination of document content and META tags.
5) Solely on document content.
6) Link and Content Analysis.

The traditional manual approach of classification would involve the analysis and classification of the contents of the web page by a number of domain experts. But this approach is inappropriate in case of Web Page Classification because of vast number of web pages available on the Internet.

Clustering algorithms have been used to form clusters of related web pages to make classification easier and faster. However, these algorithms are static because most of the clustering algorithms like K-Means etc. require the number of clusters to be specified in advance.

META tags classification techniques solely rely on content attributes of the <META name=''Keywords''> and <META name=''description''> tags of the web pages. The disadvantage with this method is that web page owners may specify keywords that are irrelevant to the contents of their web pages just to increase the hit ratio of their own pages.

The fourth and fifth methods of classification use the text content of the web page as the classification parameter. In text-based approaches, first a database of keywords in a category is prepared by counting the frequency of the keywords. The commonly occurring words (called stop words) are removed from this list. The remaining words are the keywords for that particular category and can be used for classification. To classify a document, all the stop words are removed and the remaining keywords/phrases are represented in the form of a feature vector. This document is then classified into an appropriate category using the K-Nearest Neighbor classification algorithm. These text based algorithms do not make use of other relevant features like the structure of the document, images etc. for classification.

The link-based approach is an automatic web page categorization approach based on the fact that a web page that refers to a document must contain enough hints about its content. Such hints given by the linking page can be used to classify the document being referred.

B. Feature Selection

Feature extraction or selection is an important pre-processing step in pattern recognition or pattern classification, data mining, web mining and machine learning. It also helps to remove noise features in web pages so as to improve search efficiency. The basic idea of feature selection algorithms is searching through all possible combinations of features in the data to find which subset of features works best for prediction and hence in searching. The selection is

This work was supported in part by the Departments of Computer Science and Engineering, Laksmi Narain College of Technology and Bansal Institute of Science and Technology, Bhopal, Madhya Pradesh, India.
[1] Research Scholar, Lakshmi Narain College of Technology, Bhopal
[2] Research Scholar, Bansal Institute of Science and Technology, Bhopal
[3] Asst. Professor, Bansal Institute of Science and Technology, Bhopal
[1]sinijoseph@hotmail.com, 2aishwarya.vishwakarma@gmail.com

done by reducing the number of features of the feature vectors, keeping the most meaningful ones, and removing the irrelevant or redundant features.

Feature selection algorithms designed with different evaluation criteria broadly fall into three categories:

- The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm.
- The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the mining algorithm aiming to improve mining performance, but it also tends to be more computationally expensive than the filter model.
- The hybrid model attempts to take advantage of the two models by exploiting their different evaluation criteria in different search stages.

C. Page Rank

The collection of web pages can be trated as a graph with each web page having an in-degree and an out-degree. We assume page A has pages $T_1, T_2 \ldots\ldots, T_n$ which point to it. The parameter d is a damping factor which can be set between 0 and 1with typical value of 0.85. *out_deg*(A) denotes the number of links going out of page A (out degree of A). The page rank of a page A is given as follows:

$$PR(A)=(1-d)+d\left[\sum_{i=1}^{n}\frac{PR(T_i)}{out\_deg(T_i)}\right] \quad \ldots\ldots(1)$$

Page rank or PR(A) can be calculated using a simple iterative algorithm and corresponds to the principal eigenvector of the normalized link matrix of the web. Let n be the number of documents we have . We define the link matrix M, where the M_{ij} entry is $1/n_j$ if there is a link from document j to document I, otherwise M_{ij} entry is 0. n_j is the number of the forward link of document j (out degree of j). Then we can compute the Page Rank on the graph which is the dominant eigenvector of the matrix A [3]. Thus page rank can be considered as the probability of a surfer to visit a particular page.

The rest of the paper is organized as follows: Section I enlist the various attempts made in the field of web page classification. Section II describes the proposed model. Section III shows the experimental results. Section IV emphasizes on the advantages of the proposed model and Section V concludes the paper.

II. Related Work

Chandra Chekuri, Michael H. Goldwasser, Prabhakar Raghavan, Eli Upfal (1999) propose a new architecture for web search using automatic classification[1]. In this paper a search interface is proposed that combines context-free syntactic search with context-sensitive search guided by classification. The classification process is based on term-frequency analysis and is completely statistical.

K.Selvakuberan, M. Indradevi and Dr. R. Rajaram have proposed an algorithm for combined feature selection and classification of web pages[2]. The proposed algorithm works in two stages to select initial features and final features. The experimental results show that the performance of this algorithm is better as compared to Best First, Rank Search and Forward Selection algorithms.

Lei Dong, Carolyn Watters, Jack Duffy and Michael Shepherd in their paper describe a set of experiments to examine the effect of various attributes of web genre on the automatic identification of the genre of web pages. Four different genres are used in the data set which are FAQ, News, E-Shopping and Personal Home Pages. The effects of the number of features used to represent the web pages as well as the types of attributes, <content, form, functionality>, singly and in various combinations are examined. The results indicate that fewer features produce better precision but more features produce better recall, and that attributes in combinations will always perform better than single attributes [4].

Shou-Bin Dong in his paper presents the classification of web content based on the combination of both textual and visual features. This combination is achieved by multiple classifier combination. A schema based on adaptive category weighting is proposed for achieving good combination, which has gained better results compared to the ordinary combination based on general voting schema [5].

Arkaitz Zubiaga, V´ıctor Fresno, Raquel Mart´ınez have studied Support Vector Machines which present an effective and speedy approach to solve automated classification tasks. Although SVM only handles binary and supervised problems by nature, it has been transformed into multiclass and semi-supervised approaches in several works. A previous study on supervised and semi-supervised SVM classification over binary taxonomies showed how the latter clearly outperforms the former, proving the suitability of unlabeled data for the learning phase in this kind of tasks. However, the suitability of unlabeled data for multiclass tasks using SVM has never been tested before. In their work, they present a study on whether unlabeled data could improve results for multiclass web page classification tasks using Support Vector Machines. As a conclusion, they encourage to rely only on labeled data, both for improving (or at least equaling) performance and for reducing the computational cost [6].

III. Proposed Model

Web-page classification is much more difficult than pure-text classification because of a large variety of noisy and unwanted information embedded in Web pages. It has been established that for efficient web page classification merely the textual classification is not enough. So we propose a combination approach for web page classification based on page rank and feature reduction. Web-page summaries generated by human editors can indeed improve the performance of Web-page classification algorithms. Experimental results show that the classification algorithms (NB or SVM) augmented by any summarization approach can achieve an improvement as compared to pure-text-based classification algorithms.

In our proposed model, when a web page is uploaded, the

page rank can be calculated by using equation (1) and the feature selection of the web pages can be done based on textual classification. For every web page a database is created comprising of its page rank denoted by PR(A) and its keywords $k_1, k_2 \ldots k_m$ where k_1 has the highest frequency and k_m has the least frequency. The value of *m* can be set accordingly with the consideration that:

$$m \propto \text{Time complexity}$$

The proposed model can be viewed as comprising of four phases, namely:

1) Search Phase
2) Classifier Phase
3) Sorting Phase
4) Listing Phase

The methodology is summarized by the block diagrams shown in figure 1(a) and (b) :

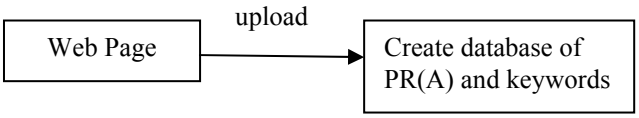

Figure 1(a)

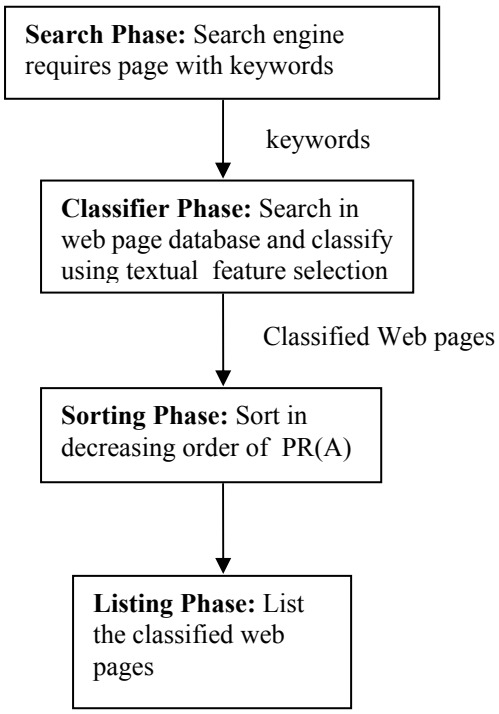

Figure 1(b)

Figure 1: Classification of web pages based on page rank and keyword frequencies

Using the feature selection method as proposed in [2] we can obtain the database with the following fields:

<keyword, number of pages classified, frequency of keywords>

Further, a new field PR(A), i.e. page rank of a web page can be computed and added to the database. Let us assume that we have *n* pages classified for a keyword and the data has index from 0 to *n*-1. We propose the algorithm for web page classification as follows:

Web page classification Algorithm

1) Set i=1
2) Repeat steps 3, 4, 5, 6 and 7 while i<n
3) Set temp=PR(A_i)
4) Set j=i-1
5) Repeat while temp<PR(A_j)
   a) PR(A_{j+1})= PR(A_j)
   b) j=j-1
   End of step 5 loop
6) PR(A_{j+1})=temp
7) i=i+1
   End of step 2 loop
8) Set i=n-1
9) Repeat while i>=0
   List page with PR(A_i)
   End of step 9 loop
10) Exit

IV. Experimental Results

In this section we are showing the implementation of our proposed algorithm on a data set of web pages. The data set used for our study contains WWW-pages collected from computer science departments of various universities in January 1997 by the World Wide Knowledge Base (Web->Kb) project of the CMU text learning group. The 8,282 pages were manually classified into the following categories:

- student
- faculty
- staff
- department
- course
- project
- other

The class other is a collection of pages that were not deemed the ``main page" representing an instance of the previous six classes. (For example, a particular faculty member may be represented by home page, a publications list and several research interests pages. Only the faculty member's home page was placed in the faculty class. The publications list and research interests pages were all placed in the other category.) The data set used in our study is available from http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/webkb-data.gtar.gz.

The files are organized into a directory structure, one directory for each class. Each of these seven directories contains 5 subdirectories, one for each of the 4 universities and one for the miscellaneous pages. These directories in turn contain the Web-pages. The file name of each page corresponds to its URL. The pages start with a MIME-header. Some of the pages do not contain useful information. For example, about 80 pages only contain information for redirecting the browser to a different location. These are not

evenly distributed over the different classes.

After the classifier phase the pages are classified and their average page rank is also calculated as shown below:

Keyword	No. of pages classified	Average PR(A)
student	1641	1.45
faculty	1124	0.73
staff	137	0.61
department	182	1.79
course	930	1.98
project	504	2.31
other	3764	2.59

TABLE 1: CLASSIFICATION OF WEB PAGES BASED ON KEYWORD FREQUENCIES WITH AVERAGE PR(A) FOR EACH KEYWORD

Let us consider a situation where the user is searching for a keyword 'student' through a search engine. The database created from the dataset will typically be in the following format:

Keyword	URL of pages classified	Page Rank PR(A_i)
student	A_0	1.73
	A_1	2.65
	A_2	1.78
	A_3	2.23
	A_4	1.39
	A_n	1.59

TABLE 2: DATABASE OF WEB PAGES WITH PAGE RANKS

Applying the proposed algorithm on the above database, we would obtain the listing of the URLs in the decreasing order of their page ranks as shown below:

Keyword	URL of pages classified	Page Rank PR(A_i)
student	A_1	2.65
	A_3	2.23
	A_2	1.78
	A_0	1.73
	A_4	1.39
	A_n	1.59

TABLE 3: DATABASE OF WEB PAGES AFTER APPLYING THE PROPOSED ALGORITHM

V. ADVANTAGES OF THE PROPOSED MODEL

Text based classification techniques solely rely on content attributes of the <META name=''Keywords''> and <META name=''description''> tags of the web pages. The disadvantage with this method is that web page owners may specify keywords that are irrelevant to the contents of their web pages just to increase the hit ratio of their own pages. This means that an author of a web page may put commonly used keywords in the <META> tags of the web pages. This can be overcome by our model because we are also taking into consideration the page rank of the web pages. Hence, malicious attempt to get a particular page listed can be overcome by our proposed approach.

Secondly, it would also yield accurate results with comparatively lesser time constraints.

VI. CONCLUSION

The World Wide Web is spreading widely and within a few years the amount of web content will surely increase tremendously. Hence, there is a great requirement to have algorithms that could classify and list web pages accurately and efficiently. In this paper we have attempted to propose a solution for web page classification using page ranks. The proposed model will provide the necessary web page classification technique for fast and efficient working of the search engines.

ACKNOWLEDGMENT

We would like to acknowledge the help rendered by the faculty and staff of Lakshmi Narain College of Technology and Bansal Institute of Science and Technology, Bhopal. The support provided for this work by our research guides is worth mentioning.

REFERENCES

[1] Chandra Chekuri, Michael H. Goldwasser, Prabhakar Raghavan, Eli Upfal (1999) "Web Search Using Automatic Classification", Sixth International World Wide Web Conference Poster Presentations

[2] K.Selvakuberan, M. Indradevi, Dr. R. Rajaram, "Combined Feature Selection and Classification-A novel approach for the categorization of Web Pages," in Journal of Information and Computing Science, Vol.3, No. 2, 2008, pp.083-089.

[3] Arun K. Pujari,"Data Mining Techniques" , Universities Press (India) Private Limited, 2003, pp.234

[4] Lei Dong, Carolyn Watters, Jack Duffy, Michael Shepherd, "An Examination of Genre Attributes for Web Page Classification", In Proceedings of the 41st Hawaii International Conference on System Sciences – 2008

[5] Shou-Bin Dong, "The hierarchical classification of Web Content by the combination of Textual and Visual Features", In Proceedings of the Third Intemational Conference on Machine Learning and Cybernetics, Shanghai, 26-29 August 2004

[6] Arkaitz Zubiaga, V´ıctor Fresno, Raquel Mart´ınez, "Is Unlabeled Data Suitable for Multiclass SVM-based Web Page Classification?, In Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing, pages 28–36,Boulder, Colorado, June 2009

[7] Loc Q. Tran, Chan W. Moon, Daniel X. Le, George R. Thoma, "Web page downloading and classification" , In Proceedings of the Fourteenth IEEE Symposium on Computer-Based Medical Systems (CBMS'01),IEEE 2001

[8] Amir Masoud Rahmani, Zahra Hossaini, Saeed Setayeshi,," Link Processing for fuzzy web pages clustering and classification", European Journal of Scientific Research,2009

[9] Lee Zhi Sam1 , Mohd Aizaini Maarof2, Ali Selamat , "Automated Web Pages Classification with Integration of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) as Feature Reduction", In Proceedings of International Conference on Man-Machine Systems, Langkawi, Malaysia, September 15-16 2006