Trading Department Oriented Web Classification Using Naïve Bayes

Nilar Htun, Moe Sanda Htun
Computer University (Sittwe)
nilarhtunsittway@gmail.com, moesdhtun@gmail.com

Abstract

Web page classification is significantly different from traditional text classification because of the presence of some additional information, provided by the HTML structure and by the presence of hyperlinks. Web classification is based on a text classification method known as Naïve Bayes. Naïve Bayes is often used in text classification applications and experiments because of its simplicity and effectiveness. In text and web page classification, Bayesian prior probabilities are usually based on term of word frequencies and term counts within a page and its linked pages. This paper presents Naïve Bayes method to classify Web pages by using keywords and defines the respective sections or departments for trading company. This paper is focused on web page representation by text content.

Keywords: Naïve Bayes, Web page classification, text classification

1. Introduction

Text classification is the assignment of predefined categories to text documents. Text classification has many applications in natural language processing tasks such as E-mail filtering, news filtering, prediction of user preferences and organization of documents [1].

Today, the fast expansion of the Web has turned Internet into a huge source of information; therefore, achieving the required information with faster speed as well as more simple, clear and efficiency way is necessary. Web Classification is a sub-field of information retrieval and web mining. Web mining is the application of data mining techniques to discover patterns from the web .web mining can be divided into three different types, which are Web usage mining, web content mining and web structure mining. Web Usage Mining is the application that uses data to analyze and discover interesting patterns of users' usage data on the web. Web content mining is the process to discover useful information from the content of a web page. Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. This paper focuses on Web content Mining [2].

When User want to find information in the Web, user usually access to it by search engines. This way to access information works well when user want to retrieve homepages, websites related to corporations, institutions or specific events, and finding quality portals [4] . Automatic web page classification can be very useful for tasks such as relating pages retrieved by search engines, or creation and maintenance of web directories. This focus on web page text content, so, hyperlinks and multimedia data are not considered. Some of the elements which can be distinguished in a web document are: plain text, text enriched with HTML tags, meta content, hyperlink structure, text associated with hyperlinks or a set of statistical parameters (media types, size, number of images or links, etc.)[6].This paper is focused on web page representation by text content. The terms or features which represent web pages will be extracted from the text of them and some information from HTML tags will be taken into account. This paper presents the web classification for sections and departments of a company by using Naïve Bayesian method. This method gets the probability that keywords appear in a department. Keywords play a major role in Naïve Bayesian classification and these words represents respective sections or departments. This paper only uses trading departments. Departments are divided into six types: Marketing, Sale & Distribution, Import/Export, Account, Finance, Human resource. Departments are classified by calculating probabilities depending on keywords.

2. Naïve Bayes Classification

Naïve Bayes classifiers are known as a simple classification algorithm and a simple probabilistic classifier based on Bayes theorem. Bayes theorem provides a method to calculate the probability of a hypothesis based on its prior probability, the probability of observing various data given the hypothesis and the observed data itself.[3] Naïve Bayes classifiers work much better in many

complex real words situation than one might expect. Naïve Bayes classifiers require a small amount of training data to estimate the parameters necessary classification. Naïve Bayes approach, which is one of the most effective approaches for text / web document classification and also a straightforward method that has proven good results in classifying documents [5].

By using Bayes' theorem,

$$P(\text{class /features}) = \frac{P(\text{class}) \times P(\text{Features/ class})}{P(\text{features})} \quad (2.1)$$

$$P(\text{class}) = \frac{\#\text{of instances in class}}{\text{total number of intances}} \quad (2.2)$$

$$P(\text{feature/class}) = \frac{\#\text{of instances with feature in class}}{\#\text{of instance in calss}} \quad (2.3)$$

$$P(\text{feature}) = \frac{\#\text{of instances in features}}{\text{total number of instances}} \times \text{total class} \quad (2.4)$$

3. Proposed System

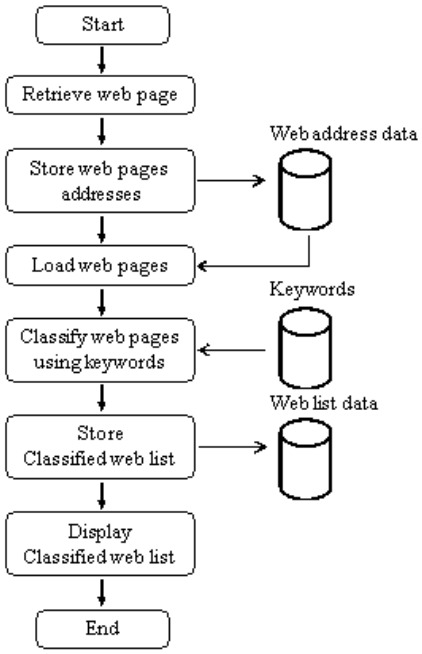

Figure 3.1: Proposed of System Design

3.1. Web Classification Algorithm

This paper evaluated different probabilities functions for trading company. We consider a well known function in automatic Web classification, Naïve Bayes probabilities, whose estimations are only based on the values of term counts within a page and among the collection. These counts are stored in database. In this algorithm, their probabilities values should be learned in this way.

Begin

Retrieve contents of web page as Text.

Find and Count the keywords occurrence stored in KEYWORD_DB (database).

Calculate the probability for each department

IF probability >= 0.3 THEN

Save the website in corresponding departments

END IF

Show the result.

End

3.2. Data Set

We created the experimental dataset which keywords that can help the system to classify departments. In this work, we fix the maximum length of keywords to be 45 characters. These keywords representing respective departments are retrieved from trading company. For simplicity and fast classification process, keywords for the departments are prepared. This paper assumes for each department used keywords in trading company. Then by using the web classification system, the collected web pages are classified and saved to output data for display on user requirement. The classification comprises of six departments:

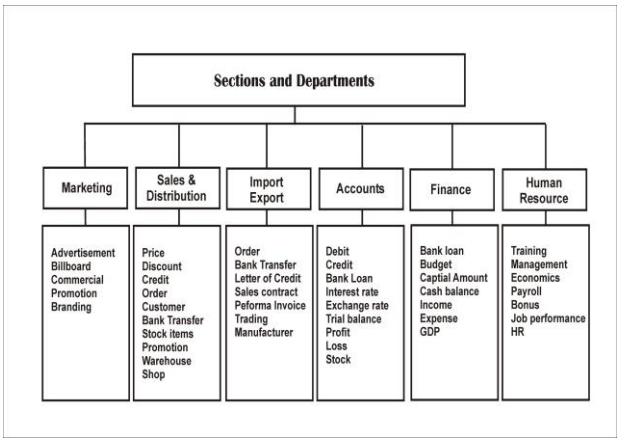

Figure 3.2: Sections or Departments associated With keywords

3.3. Web page collection

We use Web pages to evaluate possible departments of Web classification system. Therefore, we collected Web pages with these steps.

(1)We considered only the keywords that fulfilled two conditions: (i) occur more than one time in the documents and (ii) appear in more than one documents.

(2)We count the number of times each keyword appears each Web page, and the number of web pages where each keyword appears.

In the collections, the 70% of web pages were selected to classify departments, and 30% were used to evaluate the probabilities of web pages

.

Table 3.1: Sample Keywords found in Web Pages

Found Keywords		
Keywords	Department	Occurrences
Stock	Accounts	3
Price	Sales and Distribution	2
Management	Human Resource	2
Economics	Human Resource	2
Budget	Finance	3
Trading	Import/Export	3
Income	Finance	2
Loss	Accounts	2
Customer	Sales and Distribution	2

When a web page is downloaded from Internet, we count keywords that come into the web page is the frequency, which is the number of time a keyword appears in the text. By testing, we found web page's keywords for the corresponding departments. These keywords are described as follows in Table 3.1.

3.3.2. Evaluation Measures

We test possible web pages of departments in web classification algorithm. We carry out the probabilities evaluation to determine the classification results by means of the estimated with Naïve Bayes. Naïve Bayes classifier is more highly sensitive to feature selection. We present an evaluation over a large classification to test the number of keywords. For example, by using we calculate the evaluation steps for Human Resource departments.

Calculation for Human Resource Department

P (HR) = 10 / 45 = 0.2
P (Management | HR) = 2 / 4 = 0.5
P (Management) = 2 / 21*6 = 0.6
P (HR | Management) = ((0.2 * 0.5) / 0.6) = 0.16
P (HR) = 10 / 45 = 0.2
P (Economics | HR) = 2/ 4= 0.5
P (Economics) = 2 / 21*6 = 0.6
P (HR | Economics) = ((0.22 * 0.44) / 0.6) = 0.16
Answer: Probability = 0.32

3.3.3. Results

The results of experiments which one web page are summarized in Table 3.2 which shows the probabilities values after Naïve Bayes classification with keywords of different departments.

Table 3.2: Probabilities of respective departments

(Department)	**P(Keyword/Department)**	**P(keyword)**	**P(Department/Keywords)**
P(Account)=0.2	P(Stock| Account)=0.6	P(Stock)=0.8	P(Account| Stock)=0.15
P(Account)=0.2	P(Loss| Account)=0.4	P(Loss)=0.6	P(Account| Loss)=0.13
P(SD)=0.2	P(Price| SD)=0.5	P(Price)=0.6	P(SD |Price)=0.16
P(SD)=0.2	P(Customer| SD)=0.5	P(Customer)=0.6	P(SD |Customer)=0.16
P(Finance)=0.2	P(Budget| Finance)=0.6	P(Budget)=0.8	P(Finance |Budget)=0.15
P(Finance)=0.2	P(Income| Finance)=0.4	P(Income)=0.6	P(Finance |Income)=0.13
P(HR)=0.2	P(Management |HR)=0.5	P(Management)=0.6	P(HR |Management)=0.16
P(HR)=0.2	P(Economics |HR)=0.5	P(Economics)=0.6	P(HR |Economics)=0.16
P(Import/Export)=0.2	P(Trading |Import/Export)=1	P(Trading)=0.8	P(Import/Export |Trading)=0.25

By Web Classification algorithm,

Therefore, the result shows Human Resource and Sale & Distribution departments.

4. Implementation of the System

This system uses trading department. The function of the system is divided into 2 types, Admin and User. The function of Admin is to classify web page and to send classified web page to the corresponding departments. The function of user is to click web links, when web links is clicked; the original web page appeared. And user can get the required information within a short period.

When Admin is entered, admin area has two types, namely Classify files and Log off. The system must have to classify files. So, choose classify files.

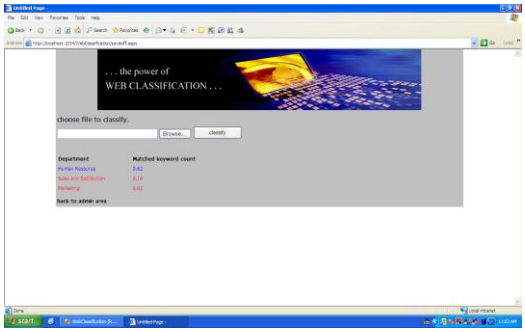

Figure 4.1: Web Classification using Naïve Bayes

Classified area has two buttons, Browse and Classify. This system chooses to classify files from Browse button and select the required web page to classify and click Classify button. This system shows probabilities of corresponding departments. Classified files will be shown with 2 colors in the corresponding department. 2 colors are blue color and red color. This system defines colors by calculating probabilities. By using Web classification algorithm, blue color is defined if probabilities have in Web page, and Admin send web page to the corresponding departments. Since user can see this web page. Red colors are defined if probabilities have not in web page, and admin does not send these Web pages. User cannot see red colors web page.

If users see blue color web page, choose back to admin area, and click logoff. and then, user can enter to get web page from login page. Users will choose the required departments.

When web link is clicked from this department, user can see and read original web page. This system can get required information from the corresponding departments.

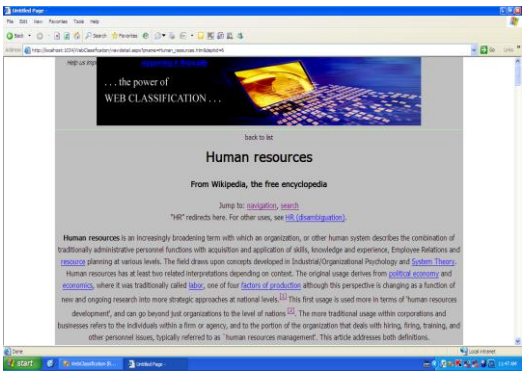

Figure 4.2: Original Web page After Classifying by Naïve Bayes

5. Conclusion

Nowadays, several approaches have been introduced to represent and classify web pages. The main approaches to web page classification have been inherited from text classification and so, only consider feature counts to carry out this task. A Naïve Bayes classification system is presented for the new enriched web page representations. Naïve Bayesian classifier will be used to classify web pages. Bayesian classifier has the minimum error rate. The goal of this system is to understand the performance of Naïve Bayes. This paper can be able to get required information from the web pages and time saving. By using this paper, user can be able information quick and easy. This paper will show to improve up to date information for Commercial Corporation and trading company.

6. References

[1] A.Mahinovs,"Text Classification Method Review"Cranfierd University, April 2007.

[2] A.Prakash Asirvatham,"Web Page Classification based on Document Structure" Internal Institute of Informational technology, Hyderabad, INDIA 500019

.

[3] I. Rish, "An empirical study of the Naïve Bayes Classifier by Workshop on Empirical Methods in Artificial Intelligence"IJCAI 2001.

[4] J.Gonzalo,"Technical report In the Software and Computing System"2004.

[5] P.Loan, "An approach of the Naïve Bayes Classifiers for the document classification" (2006), pg. 135–138.

[6] Y.Yang,"A Study of approaches to hypertext categorization" Intelligence Information system, 219,241, 2002.