ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 6, August 2012

Web Text Mining for news by Classification

Ms. Sarika Y. Pabalkar
Pad Dr. D.Y Patil Institute of Institute of Engineering and Technology, Pimpri , Pune, Maharashtra, India.

ABSTRACT— In today's world most information resources on the World Wide Web are published as HTML or XML pages and number of web pages is increasing rapidly with expansion of the web. In order to make better use of web information, technologies that can automatically re-organize and manipulate web pages are pursued such as by web information retrieval, web page classification and other web mining work. Research and application of Web text mining is an important branch in the data mining. Now people mainly use the search engine to look up Web information. The search engine like Google can hardly provide individual service according to different need of different user. However, Web text mining aims to resolve this problem. In Web text mining, the text extraction and the characteristic express of its extraction contents are the foundation of mining work, the text classification is the most important and basic mining method. Thus classification means classify each text of text set to a certain class depending on the definition of classification system. Thus, the challenge becomes not only to find all the subject occurrences, but also to filter out just those that have the desired meaning. Nowadays people usually use the search engine—Google, Yahoo etc. to browse the Web information mainly. But these search engines involve so wide range, whose intelligence level is low. It is very difficult to mine data further. The development of techniques for mining unstructured, semi-structured, and fully structured textual data has become increasingly important in industry.

Keywords— Text Mining, Extraction, Classification, Stemming, Stopword Removal

I. INTRODUCTION

The Web today contains a treasure trove of information about subjects such as people, companies, organizations, products, etc. that may be of wide interest. Web Mining is the application of data mining techniques to discover patterns from the Web. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content and usage log. Based on the primary kind of data used in the mining process, web mining can be divided in 3 categories.

- Web usage mining, is the application that uses data mining to analyze and discover interesting patterns of user's usage data on the web.
- Web content mining is the process to discover useful information from text, image, audio or video data in the web.
- Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site.

Thus Web mining is the activity of identifying term ***p*** implied in large document collection C, which can be denoted by a mapping i. e C →p [2]. A first step toward any Web-based text mining effort would be to collect a significant number of Web pages having mention of a subject. Thus, the challenge becomes not only to find all the subject occurrences, but also to filter out just those that have the desired meaning. Thus Text Mining is non trivial extraction of implicit, previously Unknown, and potentially useful information from (large amount of) textual data.

Application of text mining is:

Marketing: Discover distinct groups of potential buyers according to a user text based profile. e.g. amazon

Industry: Identifying groups of competitors WebPages .e.g., competing products and their prices.

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 6, August 2012

II. WEB TEXT MINING PROCESS

Text mining is the discovery of previously unknown information or concepts from text files by automatically extracting information from several written sources using computer software.

Text mining on Web adoptive technique include classification, clustering, association rule and sequence analysis etc.. Among them, classification is a kind of data analysis form, which can be

used to gather and describe important data set. In Web text mining, the text extraction and the characteristic express of its extraction contents are the foundation of mining work, the text classification is the most important and basic mining method. Web Text Mining Process (referring figure 1) consists of 4 steps.

III. DESIGN CONSIDERATION FOR CLASSIFICATION

The whole extraction information set can be divided into some large category. Each large category can be divided in to sub category. So, the whole extraction information set is looked as the root of the tree, and each classification as the node of the tree, the whole classification system constitutes a text classification tree.

1 *Extraction*

In extraction process, required information is extracted by checking maximum text density from the text contents from a web page. By this process, noise from the web page is removed. Extraction is followed by pre-processing of the text content. Pre-processing of the text contents include stemming and stop word removal.

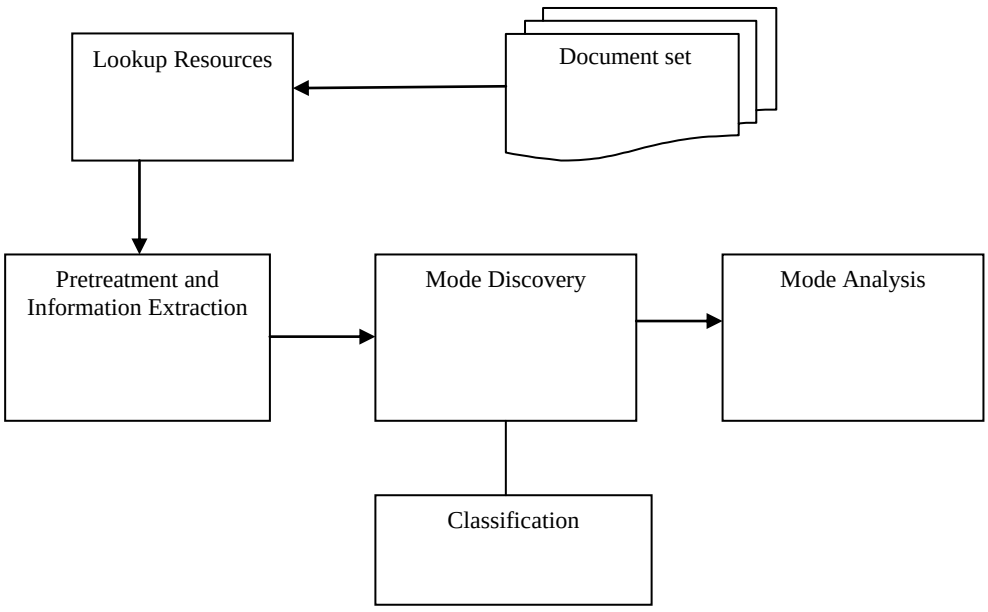

Fig 1: Web Text mining Common Process

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 6, August 2012

2 Text Classification Techniques

Vector Space Model (VSM) is adopted as a technique for classification. VSM was put forward in 60's by Salton. As the earliest and also the most famous mathematics model in information search, its basic thought is the text is assigned to authority characteristic vector. Then it confirms categories of samples that need to be divided by method of computing text the vector space model is an algebraic model used for information retrieval.

Salton's classic weighting is given by the following equation:

Term Weight

$$w_i = tf_i * log\left(\frac{D}{df_i}\right)$$

Where

- *tfi* = term frequency (term counts) or number of times a

Term *i* occurs in a document.

- *dfi* = document frequency or number of documents

Containing term *i*

- *D* = number of documents in the database

the *dfi/D* ratio is the probability of selecting a document containing a queried term from a collection of documents.

The similitude degree method of literature search technique is adopted in the system for classification, which depends characteristic vector match [1].

Suppose that the sample information is U, needed to be classified information is V, cosine of vector angle can be used to measure both of the similitude degree, it is shown as formula.

$$\therefore Co\sin e\theta_{Di} = Sim(Q, D_i)$$

$$\therefore Sim(Q, D_i) = \frac{\sum_i W_{Q,j} W_{i,j}}{\sqrt{\sum_j W_{Q,j}^2}\sqrt{\sum_i W_{i,j}^2}}$$

Classification is to classify text document in appropriate category and produce the output.

IV IMPLEMENTATION

By using HTML parser entire text data from web page is extracted excluding data containing audio, video and images. Next to fetch only required text from extracted text, text density per line is used. For getting required text , text density threshold is selected as 130 characters per line.

Pre-processing is done on extracted text by stemming and stop word removal. Stemming means to remove all the prefix and suffix from every extracted word. For stemming porter stemmer algorithm is used. Stop word removal, means removing commonly used words like is, an, or etc.

For forming a classification tree, Oracle9i database is used. Orcale9i has a feature for forming tree structure using child –parent relationship.

Text classification is a kind of typical model directive machine learning problem. It is generally divided into training and categorizing two stages. Its concrete algorithm is described as

Follows:

Training stage:

1) C={c 1, c 2, ……, cn} // Define the category set

E.G C=(Health , Country, Politics,…)

2) S ={s 1, s 2, ……, sm} // Give training text set

For i =1 to m

Training text si is marked as the sign cj that is belonged to category

End for

(3) For i =1 to m

X[si] : characteristic vector of si

X[cj]:characteristic vector is representative of each category Cj of

corresponding si.

E.g X[Software]=Java

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 6, August 2012

X[Country]=Java

End for

Categorizing stage:

(4)Define Information classification tree

Decide threshold of information and similitude degree for each leaf node

5) Add Info and Deg to classification form of corresponding X[i] sort depending on Degree & threshold

V RESULTS

This system extracts required news and classifies particular news from webpage into certain class or multiple classes and then again subs classifying it in to specific class. News is the root of the tree, it is sub classified in to four categories as Sport, Business, Nation and Education. Each category is again subcategories in to two categories. Sport is Subcategories into Cricket and Hockey/Football/Tennis. Business is subcategories into shares/mutual fund and normal business. Nation is subcategories into national and international. Education is subcategories into school and college (referring figure 5).In news is classified in education which is further classified in college and school is shown (referring figure 4).

At the same time multi classification is also implemented (referring figure 6). Multiclassification means classifying a text content in to two different categories i.e., as per the news subject context, a particular news can be categories in to two different categories.

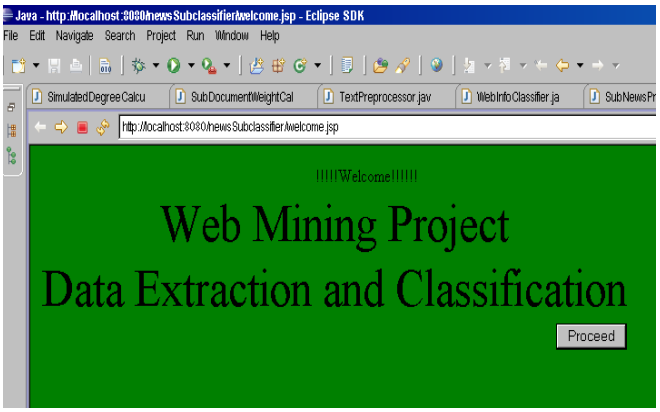

Fig 2: Screen for Home Page

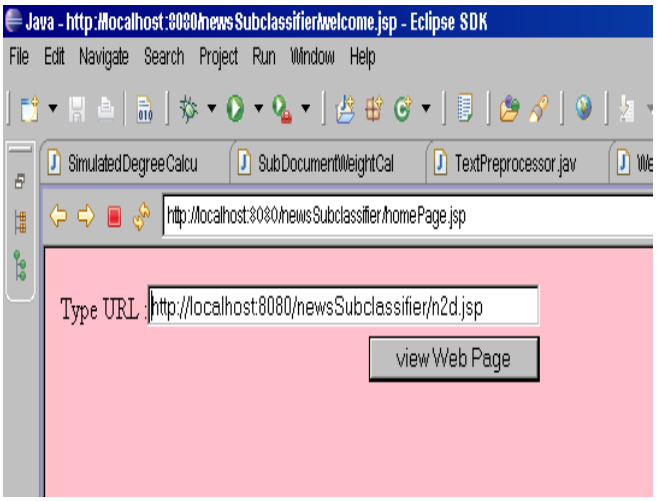

Fig 3: Screen for address of Home Page

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 6, August 2012

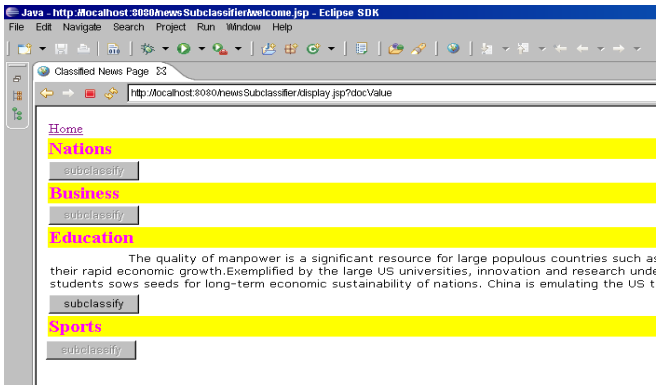

Fig 4: Screen after Classification

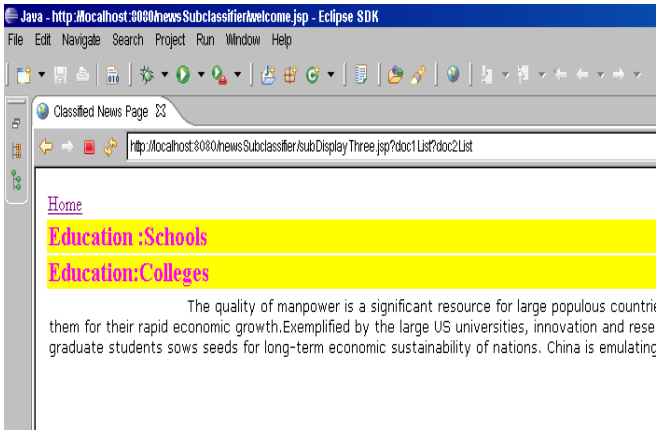

Fig 5: Screen for subclassification

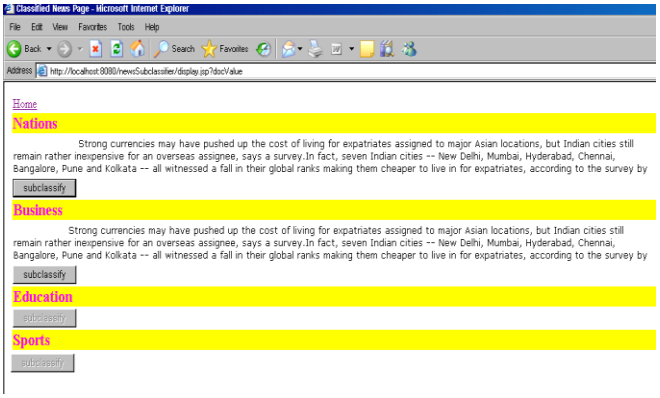

Fig 6: Screen for Muticlassification

VI CONCUSION

The Web mining classification technique is used in the system for extraction and classification. The system will provide the exact category of information to the final user. Further it can provide different personalized services according to different user's requirements [3]. It is different from search engine and intelligence technique.

References

[1.] *Shiqun Yin Gang Wang Yuhui Qiu Weiqun Zhang.Research and Implement of Classification Algorithm on Web Text Mining.IEEE2007.*

[2.] *Shiqun Yin Yuhui Qiu1,Chengwen Zhong . Web Information Extraction and Classification Method .IEEE 2007*

[3.] *Shiqun Yin Yuhui Qiu Jike Ge, Xiaohong Lan.Research and Realization of Extraction Algorithm on Web Text Mining.IEEE 2007*

[4.] *Shiqun Yin Yuhui Qiu Jike Ge..Research and Realization of Text Mining Algorithm on Web.IEEE2007*

[5.] *Wang Jicheng, Huang Yuan, Wu Gangshan and Zhang Fuyan. Web Mining: Knowledge Discovery on the Web.IEEE1999*