A Review Paper: Categorization of Web Pages

Ms. Taruna Mehta*, Ms. Payal **, Ms. Neeru Ahuja***

*** Lecturer, D.N P G College, Hisar, Haryana**

**** Student, Maharishi Markandeshwar University, Mullana, Haryana**

***** Lecturer, D.N College, Hisar, Haryana**

***er.tarunamehta@gmail.com , **payalbeniwal@gmail.com , *** neeru902@gmail.com**

Abstract

Contemporary web is comprised of trillions of pages and everyday tremendous amount of requests are made to put more web pages on the WWW. It has been difficult to manage information present on web than to create it. Web page categorization can be defined as an approach to categorize the web pages based on a set of predefined categories to manage large web content. Yahoo! and ODP are the examples of web directories in which pages are categorized manually or semi automatically, but it is a very time consuming task. There are many ways of categorizing web pages using different techniques. An approach to categorize web pages automatically on the basis of characteristics of web pages using neural network based single discrete perceptron training algorithm which is extended by selecting web page specific features to categorize web pages of predefined categories with high accuracy. The idea is presented with the help of two specific and major categories of web pages chosen for categorization that are newspaper and education.

Keywords: *Categorization, Web*

1. Introduction

Internet is the source of enormous amount of information accessed by large number of people every day. Contemporary web is comprised of trillions of pages and everyday tremendous amount of requests are made to put more web pages on the WWW. It has been difficult to manage information present on web than to create it. Web page categorization can be defined as an approach to categorize the web pages based on a set of predefined categories to manage large web content. Yahoo! and ODP are the examples of web directories in which pages are categorized manually or semi automatically, but it is a very time consuming task. There are many ways of categorizing web pages using different techniques. An approach to categorize web pages automatically on the basis of characteristics of web pages using neural network based single discrete perceptron training algorithm which is extended by selecting web page specific features to categorize web pages of predefined categories with high accuracy.

The growing number of applications on the web leads to rapid increase in number of web pages. The data available on the web can be in the form of text, images, audio, video, graphics and many other forms. Web pages present on the web can be static or dynamic. The content of dynamic web pages keeps on changing time to time. Web is considered as a large repository of information which is accessed by millions of users' everyday through internet. The dynamic nature of web and large scale explosion of web pages may put a threat to efficient information retrieval tasks. Web can be considered as an information resource, therefore it is important to describe and organize the huge content present on the web in order to realize web's full potential. Thus web page categorization is an intellectual task, important and indeed essential for organizing and understanding web content for different applications, efficient information retrieval and other tasks related to web mining. Here we will discuss some facts about web page categorization including the types of web page categorization, need of web page categorization and various characteristics of web pages.

2. Literature Review

From the very beginning categorization was done manually by domain experts. Yahoo! [3] and ODP [4] are the examples of web directories which are developed manually. But with the rapid increase of web pages it became extremely difficult to categorize web pages manually. Therefore categorization began to be done semi automatically or automatically. There are a number of approaches which have been applied in the field of web page categorization including K-Nearest Neighbor approach [11], Bayesian probabilistic models [12], inductive rule learning, decision trees, neural networks and support vector machine. All the above mentioned approaches are based only on the text content of the web pages. Besides text content other features like images, links, videos etc can also be used for categorization of web pages

the characteristics of web pages like number of links, number of images and number of words or the amount of text have been used to categorize the web pages into one of the two categories. The idea is presented using source web pages of two major categories or domains: Newspaper and Education. After analyzing the web pages belonging to newspaper sites and education sites, it has been found that newspaper web pages contain more number of links, images and words than education web pages. The difference in these characteristics is used for categorization.

3. Web Page Categorization

Web page categorization also known as web page classification is the process of assigning a web page to one or more predefined category labels. Categorization is often considered as a supervised learning problem in which a labeled data set is used to train a classifier which can be applied to classify and label the test data. The training and testing data can be collected from different sources in order to achieve high performance of the categorizer. Figure 1 shows the basic categorization method.

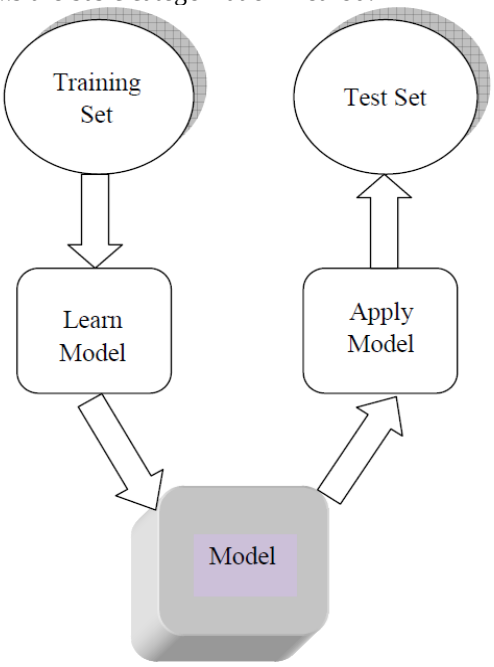

Fig 1: Basic Categorization Method

3.1 Types of Web Page Categorization

There are two types of categorizations based on the number of categories:

• **Binary Categorization:** It categorizes the web page into exactly one of two categories.

• **Multi-class Categorization:** It categorizes the web page into one of many categories.

Other types of categorizations are:

• **Subject Categorization:** It categorizes the web page according to its subject or topic. For example, categorizing the web page as "science", "sports" or "politics" is an instance of subject categorization. • **Functional Categorization:** It categorizes the web page according to its role. For example categorizing the web page as "research page", "homepage" or "information page is an instance of functional categorization.

• **Sentiment Categorization:** It categorizes the web page according to the author's attitude about any particular topic.

• **Genre Categorization:** It categorizes the web page with respect to its form or functional trait. For example when analyzing newspaper articles typical genres include "editorial", "letter", "reportage" and "spot news". On the basis of organization of categories web page categorization can also be divided into two types as explained below:

• **Flat Categorization:** In flat categorization, categories are considered parallel as shown in figure 2. The categories like "business", "sports", "health" forms a flat categorization because no category can supersede the other category.

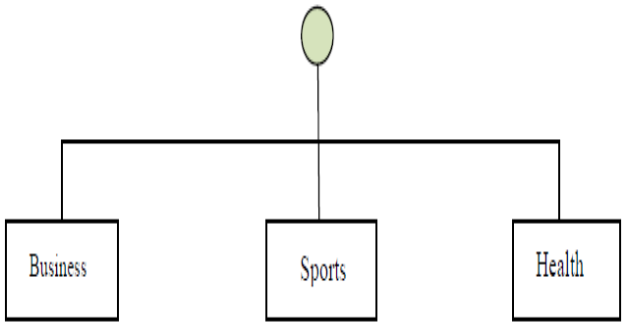

Fig 2: Flat Categorization

• **Hierarchical Categorization:** In hierarchical categorization one category can supersede the other categories as shown in figure 3.

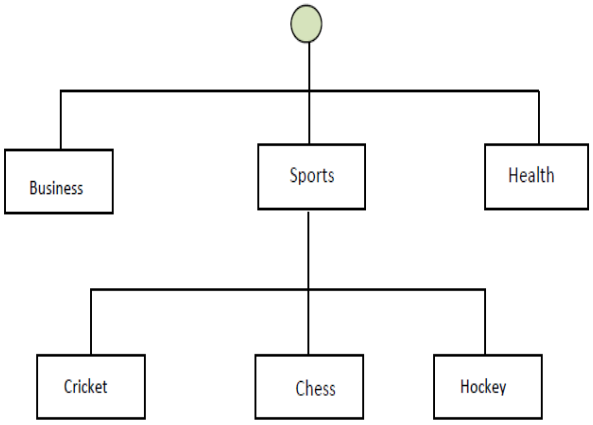

Fig 3: Hierarchical Categorization

3.2 Need of Web Page Categorization

Web page categorization is needed due to the following reasons:

• It helps in efficient retrieval of web pages.

• It provides an aid to topical crawlers which search the web for a particular topic.

• It helps in maintenance and development of web directories.

• It helps in topic specific web link analysis.

• It helps in increasing the quality of search results. Categorized results present a good user interface than the search results which are presented in a ranked list.

3.3 Characteristics of Web Page

A web page has the following characteristics:

• It is a semi structured document in HTML.

• It consists of text, images, links, videos and other multimedia content.

• It is connected to other pages through hyperlinks thus forming a graphical structure on the web.

• It is rendered to user by the web browser.

4. Web Page Categorization Techniques

Web page categorization is a fundamental problem these days due to rapid increase in the number of web pages. The need for automated categorization of web pages is for at least two reasons. One reason is the large number of resources present on the web and their ever-changing nature. It is not possible to manage such dynamic nature of web manually without a lot of human effort and time. The second reason is that categorization itself is a subjective activity; different applications depend upon different classification schemes. Therefore different types of categorization schemes, representing different facets of knowledge may need to be applied in an ongoing fashion due to large scale increase in applications [1]. A number of techniques have been used for the categorization of web pages based on different approaches as described below.

The categorization techniques can be classified into the following broad categories:

• Categorization by domain experts

• Clustering approaches

• Meta tags based approach

• Text content based categorization

• Link and Content Analysis

In manual categorization approach, categorization is done by domain experts. However it is a very time consuming task and it takes a lot of human effort to categorize the large number of web pages.

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 based approach relies on the use of meta tags in web pages like <META name=”keywords”> and <META name=”description”>. However this approach fails in the cases where web pages don’t contain meta tags. In text content based categorization, a database of keywords is prepared by calculating the frequency of occurrence of words and phrases in a category. The commonly occurring words like “the”, “and”, “of” etc. are removed from database and the remaining keywords are then used for categorization. The link and content analysis is based on the hyperlinks and anchor text present on the web page which gives enough hints about referred page.

Every web page categorization technique involves the following steps for web page categorization:

Step 1: Understand completely the domain to be categorized.

Step 2: Collect training data for the categorization.

Step3: Pre-process data by reducing the dimensions of feature set as required by the categorization algorithm.

Step 4: Put the categorizer on training.

Step 5: Apply the test data to the categorizer.

Step 6: Evaluate the results.

5. Conclusion

Web page categorization is one of the challenging tasks due to ever increasing traffic of web pages. A number of researches have been done in this field using different approaches and techniques as described above. Each one of them has some limitations. Web pages are connected to each other by hyperlinks. Feature extraction is considered as the most important task of web page categorization and also the difficult one due to semi–structured source code and hyperlinked structure of web pages. Features can be divided into two: on page features and neighboring features. On page features are the features which can be directly extracted from the web page through textual content, visual content and various HTML tags present in web pages. Neighboring features are the features that can be extracted from the web pages which are connected to web page which is needed to be categorized. But it is very difficult to extract these features. Most of the algorithms rely only on the text content of the web pages and also difficult to implement. However besides text, each type of web has its own layout. The characteristics of web pages can also be used to categorize web pages. To conclude we can say that the technique which can categorize the web pages based on some characteristics of web pages which is easy to understand and use.

6. Future Scope

• To study and analyze different features of source web pages and select those features on the basis of which web pages can be categorized.

• To build a binary categorizer and train it with input values which consist of features extracted from web pages.

• To test the binary categorizer by comparing actual output and the desired output.

• To verify and analyze the result in support of this proposal.

7. References

[1] Pierre J. M., "Practical Issues for Automated Categorization of Web Pages," September 2000.

[2] Xiaoguang Q. and Davison B. D., "Web page classification: Features and algorithms," ACM Computing Surveys, 41(2), 2009

[3] Yahoo!, http://www.yahoo.com, Accessed date 14th March, 2012.

[4] Open Directory Project, http://www.dmoz.org, Accessed date 15th March, 2012

[5] Xu Z. et. al., "A Web Page Classification Algorithm Based On Link Information," in DCABES'11 Proceedings of the Tenth International Symposium on Distributed Computing and Applications to Business, Engineering and Science , pp. 82-86, 2011.

[6] Bartik V., "Text-Based Web Page Classification with Use of Visual Information," in ASONAM'10 Proceedings of the International Conference on Advances in Social Network Analysis and Mining, pp. 416-420, 2010.

[7] He Z. and Liu Z., "A Novel Approach to Naïve Bayes Web Page Automatic Classification," in FSKD'08 Proceedings of the Fifth International Conference on Fuzzy System and Knowledge Discovery, pp. 361-365, 2008.

[8] Radovanović M. and Ivanović M.,"Document Representation for Classification of Short Web Page Descriptions," in Yugoslav Journal of Operations Research, I8, Number 1, pp. 123-138, 2008.

[9] Dai W. et. al., "A Novel Web Page Categorization Algorithm Based on Block Propagation Using Query-Log Information," in WAIM'06, LNCS 4016, pp. 435-446, 2006.

[10] Materna J., "Automatic Web Page Classification," in RASLAN'08 Proceedings of Recent Advances in Slavonic Natural Language Processing, pp. 84-93, 2008. Page | 38

[11] Kwon O. and Lee J., "Web page classification based Nearest Neighbor approach," in IRAL'00 Proceedings of the fifth international workshop on Information retrieval with Asian languages, pp. 9-15, 2000.

[12] McCallum A. and Nigam K., "A Comparison of Event Models for Naive Bayes Text Classification,"