# Web Page Categorization To Enhance Searching

Lalit kumar
amity university
Gurgaon,Haryana(India)

Pooja Batra
Amity University
Gurgaon,Haryana(India)

Asha Sohal
Amity University
Gurgaon Haryana(India)

**Abstract**

We can define Web page categorization as an approach to categorize the unorganized web pages based on a set of already defined categories to manage large web content of various new applications. There are many ways to categorize web pages using different techniques like based on extracted text features of web pages. Here we present 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. The approach can be effectively used to categorize web pages into broad categories. The whole approach can be described in three steps. In the first step, features are extracted automatically after analyzing the source web pages. The second step includes the implementation and training of the algorithm. Finally, the third step will categorize the source web pages into one of the two categories.

Keywords: *Categorization, Web*

## 1. Introduction

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. 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. There are many ways of categorizing web pages using different techniques. 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. Figure 1 shows the basic categorization method.



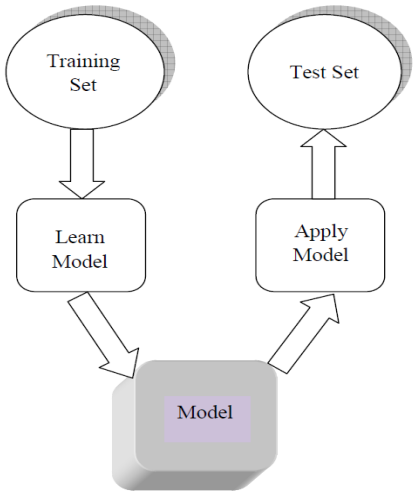


Fig 1: Basic Categorization Method

### 1.1 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.

1.2 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

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.

## 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 machines. 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. Problem Definition

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. Thus the problem is to implement the technique which can categorize the web pages based on some characteristics of web pages which is easy to understand and use.

## 4. Objective

The main objectives that are addressed in this paper is to solve the above mentioned problem are as follows:

• 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.

## 5. Methodology Used

• Collection of data set.

• Study and analyzing of the dataset.

• Selection and extraction of features from the data set.

• Implementation and training of the algorithm.

• Verification and analyzing the categorizer using test data set.

• Performance Evaluation.

## 6. Implementation

The proposed approach is explained in the following steps:

### 6.1 Data Set Collection

First of all data is collected. Data set consists of education and newspaper web pages. These web pages are collected from different sites and also from Yahoo! [3] web directory.

### 6.2 Feature Extraction

After collecting the data set, features of the web pages in data set are extracted automatically. The main features which are extracted are number of links, number of images and amount of text present on the web pages. After analyzing these features it has been found that the newspaper web pages contain more number of links, images and words as compared to education web pages. It helps in differentiating these two types of web pages.

After analyzing the values obtained for different extracted features, mean and standard deviation is calculated and each value is mapped to the value in the range [-2,2] as shown below:

| No. of Images | Input Value |
|---|---|
| 1-30 | -2 |
| 31-60 | -1 |
| 61-90 | 0 |
| 91-120 | 1 |
| 121-150 | 2 |

Table1: Input Values for Number of Images

| No. of Links | Input Value |
|---|---|
| 1-100 | -2 |
| 101-200 | -1 |
| 201-300 | 0 |
| 301-400 | 1 |
| 401-500 | 2 |

Table 2: Input Values for Number of Links

| No. of Words | Input Value |
|---|---|
| 1-1000 | -2 |
| 1000-2000 | -1 |
| 2000-3000 | 0 |
| 3000-4000 | 1 |
| 4000-5000 | 2 |

Table 3: Input Values for Number of Words

### 6.3 Training Of Algorithm

The discrete perceptron training algorithm is used in the proposed approach. It is based on the concept of neural networks. The implementation of the algorithm is done in TurboC2 using object oriented programming language C++. The pseudo code is listed in appendix. The platform used is Intel 64-bit with Core 2 Duo processor having a frequency of 2.0 GHz with Windows 7 64-bit Enterprise Edition running on it. The system had a RAM of 2.0 GB. The first step after implementation to build a categorizer

is to train it with the collected data set. The single perceptron training algorithm is based on perceptron learning rule.

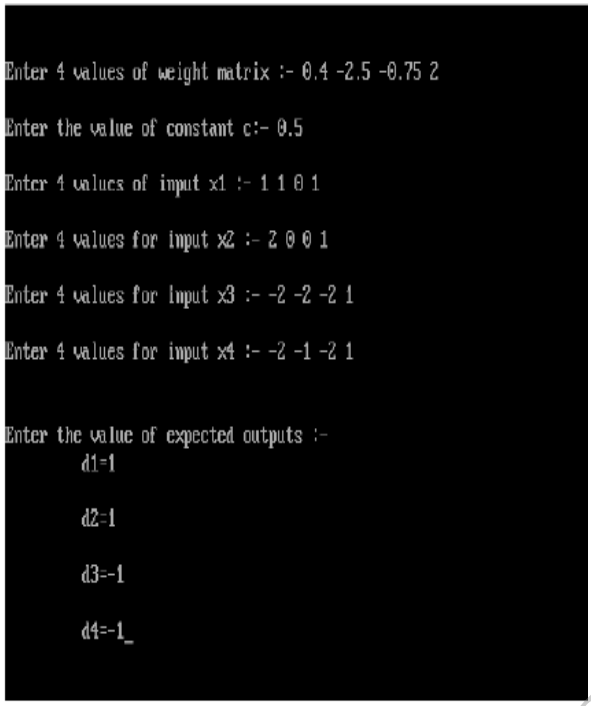


Fig 2: Training of algorithm


Fig 3: Final Weights after Training

6.4 Categorization of web pages

Once the weights are fixed, training will get completed. Testing data set can be applied to the program to categorize the web pages. 120 web pages are used to test the categorizer.

**7. Result**

Out of 120 source web pages 110 web pages are categorized correctly. Accuracy of the results can be measured in terms of precision which can be defined as the number of correct categories assigned divided by the total number of categories assigned. The experimental or testing results are shown in table 4 and 5 along with accuracy.

| | |
|---|---|
| Total Pages | 60 |
| Right Categorized Pages | 58 |
| Wrong Categorized Pages | 2 |
| Accuracy | 96% |

Table 4: Experimental Results for Education Web Pages

| | |
|---|---|
| Total Pages | 60 |
| Right Categorized Pages | 52 |
| Wrong Categorized Pages | 8 |
| Accuracy | 86.66% |

Table5: Experimental Results for Newspaper Web Pages

Hence the average accuracy obtained in the results is 91.33 percent which is very high. Figure 4 and 5 are depicting net values calculated corresponding to test input vectors along with their categories.

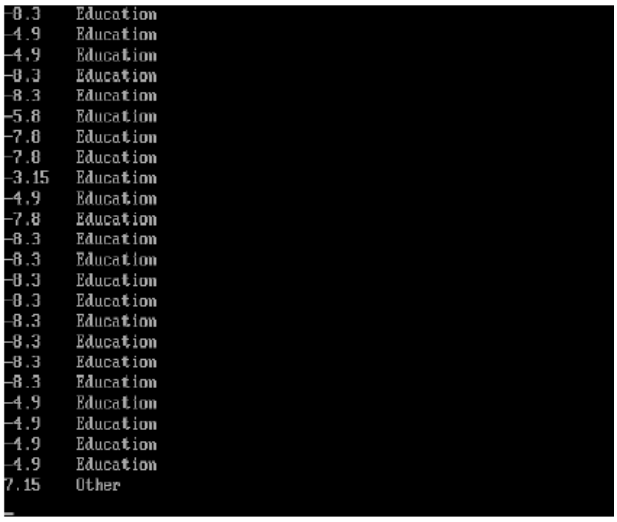


Fig 4: Categorization Output for Education


Fig 5: Categorization Output for Newspaper

## 8. Conclusion

In support of this proposal the above mentioned characteristics are extracted from newspaper and education web pages. It has been found that newspaper web pages have more number of links, images and words than education web pages. This difference helped in differentiating between the two categories. The binary categorizer built for the categorization of web pages is based on the concept of neural networks. Neural networks can be used as categorizers. The algorithm used is single discrete perceptron training algorithm. It is implemented and trained with finite set of input data vectors. After training of the algorithm, final weights are obtained which can't be modified later by any number of input data vectors. Testing is performed with 120 home pages of different newspaper and education web sites and it is found that the results obtained from this approach are 91.33 percent accurate. It can also be used to categorize web pages into broad categories. For example in order to classify blog and non blog sites, extract those features which can distinguish between the two categories. In blog sites one can find a lot of text in the form of articles, comments with lots of emoticons and also number of links. Such features can be extracted and used for categorization. Similarly one can distinguish between the research pages and content pages by analyzing the characteristics of web pages. Social networking sites and non social networking sites can also be categorized by analyzing the features which can distinguish between their characteristics.

## 9. Future Scope

Knowledge discovery which can be applied to various applications takes the services of web page categorization. Leaving behind current Categorization methods more accurate methods of categorization of pages can be developed by using various other features of web pages. It is because visual & other multimedia features still have not been used in large scale. More & more research can be done relating to algorithm's modifications. Using the Meta Tags is a good options but since most of pages don't have meta tags so automatic methods of these meta tags creation needed to be developed first and thus makes good scope for future research work.



## 10. 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," in AAAI-98 Workshop on Learning for Text Categorization, 1998.