

Web Page Categorization based on Document Structure

Arul Prakash Asirvatham
arul@gdit.iiit.net

Kranthi Kumar. Ravi
kranthi@gdit.iiit.net

Centre for Visual Information Technology
 International Institute of Information Technology
 Gachibowli, Hyderabad, INDIA 500019

Abstract

The web is a huge repository of information and there is a need for categorizing web documents to facilitate the indexing, search and retrieval of pages. Existing algorithms rely solely on the text content of the web pages for categorization. However, web documents have a lot of information contained in their structure, images, audio video etc present in them. In this paper, we propose a method for automatic categorization of web pages into a few broad categories based on the structure of the web document and the images present in it.

1. Introduction

There is an exponential increase in the amount of data available on the web recently. According to [1], the number of pages available on the web is around 1 billion with almost another 1.5 million are being added daily. This enormous amount of data in addition to the interactive and content-rich nature of the web has made it very popular. However, these pages vary to a great extent in both the information content and quality. Moreover, the organization of these pages does not allow for easy search. So an efficient and accurate method for classifying this huge amount of data is very essential if the web is to be exploited to its full potential. This has been felt for a long time and many approaches have been tried to solve this problem.

To start with, classification was done manually by domain experts. But very soon, the classification began to be done semi-automatically or automatically. Some of the approaches used include text-categorization based on statistical and machine-learning algorithms [3], K-Nearest Neighbor approach

[2], Bayesian probabilistic models [4][6][7], inductive rule learning [5], decision trees [7], neural networks [8] and support vector machines [9]. An effort was made to classify web content based on hierarchical structure [10].

However, besides the text content of the web page, the images, video and other multimedia content together with the structure of the document also provide a lot of information aiding in the classification of a page. Existing classification algorithms, which rely solely on text content for classification, are not exploiting these features. We have come up with a new approach for automatic web-page classification based on these features.

The paper is organized as follows. In Section 2, we discuss the existing algorithms for web page classification and the drawbacks in these algorithms. We discuss our approach in Section 3 followed by our implementation of this approach in Section 4. Our results are dealt with in Section 5. We present our conclusions and avenues for future work in this direction in Section 6.

2. Web Page Categorization Algorithms

Several attempts have been made to categorize the web pages with varying degree of success. The major classifications can be classified into the following broad categories

a) Manual categorization: The traditional manual approach to classification involved the analysis of the contents of the web page by a number of domain experts and the classification was based on the textual content as is done to some extent by Yahoo [11]. The sheer volume of data on the web rules out this approach. Moreover, such a classification would be subjective and hence open to question.

b) Clustering Approaches: Clustering algorithms have been used widely as the clusters can be formed directly without any background information. However, most of the clustering algorithms like K-Means etc. require the number of clusters to be specified in advance. Moreover, clustering approaches are computationally more expensive.

c) META tags based categorization: These classification techniques rely solely on attributes of the meta tags (<META name='keywords'> and <META name='description'>). However, there is a possibility of the web page author to include keywords, which do not reflect the content of the page, just to increase the hit-rate of his page in search engine results.

d) Text content based categorization: The fourth and fifth approaches use the text content of the web page for classification. In text-based approaches, first a database of keywords in a category is prepared as follows - the frequency of the occurrence of words, phrases etc in a category is computed from an existing corpus (a large collection of text). 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 as a feature vector. This document is then classified into an appropriate category using the K-Nearest Neighbor classification algorithm.

These approaches rely on a number of high quality training documents for accurate classification. However, the contents of web pages vary greatly in quality as well as quantity. It has been observed [12] that 94.65% of the web pages contain less than 500 distinct words. Also the average word frequency of almost all documents is less than 2.0, which means that most of the words in a web document will rarely appear more than 2 times. Hence the traditional method based on keyword frequency analysis cannot be used for web documents.

e) Link and Content Analysis: The link-based approach is an automatic web page categorization technique based on the fact that a web page that refers to a document must contain enough hints about its content to induce someone to read it. Such hints can be used to

classify the document being referred as has been done according to [13]

We observe that the methods used so far, are based to a great extent on the textual information contained in the page. We now present our approach based on structure of the page and image and multimedia content in the page.

3. Structure based approach

Structure based approach exploits the fact that there are many other features apart from text content, which form the whole gamut of information present in a web document. A human mind categorizes pages into a few broad categories at first sight without knowing the exact content of the page. It uses other features like the structure of the page, the images, links contained in the page, their placement etc for classification.

Web pages belonging to a particular category have some similarity in their structure. This general structure of web pages can be deduced from the placement of links, text and images (including equations and graphs). This information can be easily extracted from a html document.

We observe that there are some features, like link text to normal text ratio which are present in all categories of documents but at varying degrees and there are some features which are particular to only some kinds of documents. Hence, for the classification we have assumed apriori that a certain feature contributes to the classification of the page into a particular category by, say some $x\%$ whereas the same feature contributes to classification of the same page into some other category by some $(x \pm y)\%$. These apriori values will then be modified based on learning from a set of sample pages. The final weights are then for categorization.

4. A Specific Implementation

We have implemented a structure based categorization system to categorize the web pages into the following broad categories.

1. Information Pages.
2. Research Pages.
3. Personal Home Pages.

A typical information page has a logo on the top followed by a navigation bar linking the page to other important pages. We have observed that the ratio of link text (amount of text with links) to normal text also tends to be relatively high in these kinds of pages.

Research pages generally contain huge amounts of text, equations and graphs in the form of images etc. The presence of equations and graphs can be detected by processing the histogram of the images.

Personal home pages also tend to have a common layout. The name, address and photo of the person appear prominently at the top of the page. Also, towards the bottom of the page, the person provides links to his publications if there are any and other useful references or links to his favorite destinations on the web.

Our implementation takes a start page and a domain as input, spiders the web retrieving each of the pages along with images. The retrieval is breadth-first and care has been taken to avoid cyclic search. The pages and images are stored on the local machine for local processing and feature extraction.

The categorization is carried out in two phases. The first phase is the feature extraction phase and the second, classification phase. These are explained in detail in the following sub-sections.

4.1 Feature Extraction

The feature extraction phase is the most important phase in the system as the classification is based on the features extracted during this phase. The features should be should provide some valuable information about the document and at the same time be computationally inexpensive. Very small images (approximately 20x20 or less) have not been considered for feature extraction as they usually correspond to buttons, icons etc.

The following features were taken into consideration for classification of the pages.

(a) Textual Information: The number and placement of links in a page provides valuable information about the broad category the page belongs to. We have computed the ratio of number of characters in links to the total number

of characters in the page. A high ratio means the probability of the page being an information page is high. Some of the information pages contain links followed by a brief description of the document referred to by the link. In such cases, this ratio turns out to be low. So, the placement of the links is also an important parameter.

The amount of text in a page gives an indication of the type of page. Generally, information and personal home pages are sparse in text compared to research pages. Hence we have counted the number of characters in the document (after removing some of the commonly used words) and used this information to grade the text as sparse, moderate or high in information content.

(b) Image Information: The number of distinct colours in the images has been computed. Since there is no need of taking into account all 65 million colour shades in a true colour image, we have considered only the higher order 4 bits of each colour shade (to give a total of 4096 colours). Information pages have more colours than personal home pages, which in turn have more colours than research pages.

The histogram of images has been used to differentiate between natural and synthetic images. The histogram of synthetic images generally tends to concentrate at a few bands of colour shades. In contrast, the histogram of natural images is spread over a larger area. Information pages usually contain many natural images, while research pages contain a number of synthetic images representing graphs, mathematical equations etc.

The research pages usually contain binary images and further information can be extracted from them. Graphs are generally found in research pages. We have used the histogram of the image to detect the presence of a graph in the image.

4.2 Classification

We have used the following approach to classify the page according to the features obtained. In this algorithm, each feature contributes either positively or negatively

towards a page being classified to a particular category. The actual procedure is as follows: Let W be a $(c \times n)$ weight matrix and F be a $(n \times 1)$ feature value matrix.

$V = W \times F$ gives a $c \times 1$ matrix as shown below:

$$W = \begin{bmatrix} W_{00} & W_{01} & - & - & W_{0c} \\ W_{10} & W_{11} & - & - & W_{1c} \\ - & - & - & - & - \\ - & - & W_{ij} & - & - \\ W_{n0} & W_{n1} & - & - & W_{nc} \end{bmatrix},$$

$$F = \begin{bmatrix} F_{00} \\ F_{01} \\ - \\ F_{0i} \\ - \\ F_{0n} \end{bmatrix}, \quad V = \begin{bmatrix} V_{00} \\ V_{01} \\ - \\ V_{0i} \\ - \\ V_{0c} \end{bmatrix}$$

n – Number of features

In the matrix W ,

c - Number of Categories

and $W_{ij} \in [-1, +1]$ is the contribution of the i^{th} feature to the j^{th} category.

In the matrix F ,

F_{0i} is the value of the i^{th} feature.

For each category, W_{ij} varies between -1.0 and 1.0. For example, in case of a home page, a weight of 1.0 could be assigned for the feature ‘presence of image with human face’ whereas the same will have a weight of -0.8 in a research page. Thus the presence of a photograph increases the chance of it being classified as a personal home page and at the same time decreases the chance of it being classified as a research page or an information page.

Initially, these weights were assigned based on heuristics. These weights were later modified based on the implementation runs on sample pages. The accuracy of the output is checked, and the error is used as a feedback to modify the weights.

The classification is done based on the elements of matrix V obtained by the above procedure. The document belongs to the category for which the value of V_{ij} is the highest.

5. Results

We tested our implementation of this approach on a sample space of about 4000 pages. These pages are gathered from various domains. The results and various parameters used were shown in Table 1.

Details	Results
No. of pages on which we have tested our implementation	~4000
Pages categorized	~3700
Pages categorized correctly	~3250
% categorized correctly	87.83%

Table 1

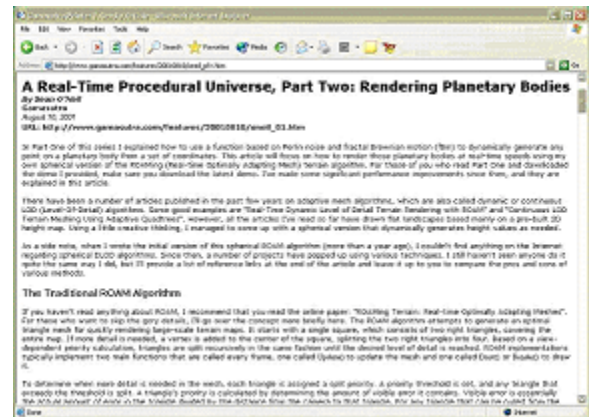


Figure 1: wrongly categorized

We have manually tested the results obtained by our implementation and found some results which are out of sync with the actual category the page should belong to. For example, the page shown in figure 1 fell into a research page though it should have fallen into information page. The reason can be attributed to the fact that it is a print version of the original

page and hence there is not a single link in the page. Also, no images are present in the page and the amount of text is very large.

6. Conclusions and Future Work

We have described an approach for the automatic categorization of web pages that uses the images and the structure of the page for classification. The results obtained are quite encouraging. This approach, augmented with traditional text based approaches could be used by search engines for effective categorization of web pages.

Our current implementation uses a method for categorization, wherein the weights assigned to each feature are set manually during the initial training phase. A neural network based could be employed to automate the training process.

Adding a few more features based on heuristics, (e.g. the classification of a page as a home page by detecting a face at the top) would increase the classification accuracy.

Currently, we have used only images in addition to the structural information present in web pages. We are not exploiting other information present in the form of video, audio etc. These could also be used to get valuable information about the page.

We have currently used our approach to categorize the web pages into very broad categories. The same algorithm could also be used to classify the pages into more specific categories by changing the feature set.

7. References

[1] John.M.Pierre, *Practical Issues for Automated Categorization of Web Pages*, September 2000.
 [2] Yiming Yang, Xin Lui *A Reexamination of Text Categorization methods*, In proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99), pp. 42-49, University of California, Berkeley, USA 1999.

[3] Oh-Woog Kwon, Jong-Hyoek Lee, *Web page classification based on k-Nearest Neighbor approach*
 [4] Andrew McCallum and Kamal Nigam, *A Comparison of Event Models for Naïve Bayes Text Classification*, In AAAI-98 Workshop on Learning for Text Categorization, 1998
 [5] Chidanand Apte and Fred Damerau, *Automated Learning of Decision rules for Text Categorization*, ACM Transactions on Information Systems, Vol 12, No.3, pp.233-251, 1994.
 [6] Koller, D. and Sahami, M, *Hierarchically classifying documents using very few words*, Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), pp.170-178, 1997.
 [7] Lewis, D.D. and Ringuette, M. A *Classification of two learning algorithms for text categorization*, Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94), pp.81-93, 1994.
 [8] Weigend, A.S., Weiner, E.D. and Peterson, J.O., *Exploiting Hierarchy on Text Categorization*, Information Retrieval, I(3), pp.193-216, 1999.
 [9] Dumais, S.T., Platt, J., Heckerman, D., and Sahami, M., *Inductive Learning Algorithms and representations for text categorization*, Proceedings of the Seventh International conference on Information and Knowledge Management (CIKM'98), pp.148-155, 1998.
 [10] Susan Dumais, Hao Chen, *Hierarchical Classification of web content*
 [11] <http://www.yahoo.com>
 [12] Wai-Chiu Wong, Ada Wai-Chee Fu, *Incremental Document Clustering for Web Page Classification*, Chinese University of Hong Kong, July 2000.
 [13] Guiseppe Attardi, Antonio Gulli, Fabrizio Sebastiani, *Automatic Web Page Categorization by Link and Context Analysis*.