Web Page Classification based on Document Structure

Arul Prakash Asirvatham
arul@gdit.iiit.net

Kranthi Kumar. Ravi
kranthi@gdit.iiit.net

International Institute of Information Technology
Hyderabad, INDIA 500019

Abstract

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

Keywords: Web page categorization, structure based categorization, document categorization, weighted feature categorization.

1. Introduction

There is an exponential increase in the amount of data available on the recently. According to [1], the number of pages available on the web is around 1 billion with almost another 1.5 million added per day. This enormous amount of data in addition to the interactive and content-rich nature of the web has made the web 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 an easy search. Hence it becomes very difficult for a person to find the document he wants related to particular information he is searching for. 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 were taken to solve this problem.

Earlier, domain experts did this classification manually. But very soon, the classification had to be done semi-automatically or automatically. Some of the

approaches according to [2] are text-categorization based statistical and machine-learning algorithms like K-Nearest Neighbor approach [3], Baynesian probabilistic models [4][6][7], inductive rule learning [5], support vector mechanics[9], neural networks[8] and decision trees[7]. Very few learning methods exploit the hierarchical structure and an effort was made by [10] to classify web content based on hierarchical structure for classification.

Besides the text content of the web page, the images, video and other multimedia content and 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.

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. We have come up with an innovative idea for automatic web-page classification based on this approach.

The paper is organized as follows. In Section 2, we discuss the existing algorithms for web page classification and the inconsistencies 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 Classification 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

1. Manual classification by domain specific experts.
2. Clustering approaches.
3. META tags (which server the purpose of document indexing).
4. A combination of document content and META tags.
5. Solely on document content.
6. Link and Content Analysis.

Manual Classification: The traditional manual approach to classification would involve the analysis of the contents of the web page by a number of domain experts and

classification based on the textual content as 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. This resulted in efforts to automate the entire classification process. However, this is based on a number of positive and negative training sets, for which again a number of domain experts are required.

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.

META tags: These classification techniques solely rely on content attributes of the <META name=''Keywords''> and <META name=''description''> tags. Though relying on these tags might give accurate results to a large extent, there is a possibility of the web page author to include keywords that don't reflect the content of the page, just to increase the hit-rate of his page in search engine results. So, some search engines that relied on this method failed to appropriately classify the web documents.

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 amount 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 in the form of 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, as mentioned earlier that 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. Moreover, these text based algorithms do not make use of other relevant features like the structure of the document, images etc for classification.

Hence the traditional method based on keyword frequency analysis cannot be used for web documents.

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 [13]. 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 relies on the fact that there are many other features apart from text content which form the whole gamut of information present in a web document. Structure based approach tends to exploit this fact.

Web pages belonging to a particular category have some similarity in their structure. Based on these similarities, any web page can be categorized into at least three 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.

In contrast, personal home pages also tend to have a common layout. The name and address of the person appear prominently at the top of the page. Included also is a photograph of the person concerned generally. 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. The presence of the photograph of a person can be detected by subjecting the images in the page to face-detection algorithms. This

augmented with the layout of the page and the placement of the links helps in classification of a page into this category.

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. There are also algorithms to detect the presence of graphs and equations in an image. The number of distinctive gray levels/color shades in the images also provides a cue about the page type. For example, synthetic images contain very few colors compared to a natural image, which contains a very large number of colors.

The general structural information of any page 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.

As far as the classification is concerned, we see 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 that are particular to only some kinds of documents.

So, for the classification we have an apriori approach where we assume 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±y)%. These apriori values will then be modified based on semi-automatic learning from a set of sample pages. The final values will then be used to categorize web pages.

4. A Specific Implementation

We have implemented a system that uses a structure-based classification approach to classify the web pages into three broad categories. We tested this program on a sample space of nearly 4000 pages belonging to various universities and other pages on the web gathered from different domains.

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. These 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. We have used a set of features that try to capture the structure of the page by analyzing the placement of text, links, images etc and deduce the contents of the page by analyzing the images in the page. Very small images (approximately 20x20 or less) have not been considered for feature extraction as they usually correspond to buttons, icons etc.

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 and used this information to grade the text as sparse, moderate or high in information content at the same time taking into account the commonly used words. The following featues 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 but still the page generally tends to be an information page. So, the placement of the links also tends to be an important parameter.

(b) Image Information: The images in the page also give an indication of the page content. Generally, information pages are more ‘colorful’ than research or home pages

i.e. the number of distinct colors is more in the case of information pages. We have extracted a few features to give an indication of this 'colorfulness'. We have also used the histogram of images to distinguish between them.

The number of distinct colours in the images has been computed. Since there is no need for 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). As mentioned earlier, information pages have more colours than personal home pages, which in turn have more colours than research pages. The research pages usually contain binary images and further information can be extracted from them. 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. This fact has been used to differentiate between natural and synthetic images. Information pages usually contain many natural images, while research pages contain a number of synthetic images representing graphs, mathematical equations etc.

Mathematical equations are generally found in research pages. We have used a standard implementation to detect the presence of mathematical equations, graphs etc in an image.

(c) Other Information: Though the approach suggested by us includes classification based on video and other multimedia content etc, our current implementation does not take this kind of information into consideration. This will be dealt with in the future implementations.

4.2 Classification

Once we get the features, we need to classify the pages based on these features. We have used the following approach to classify the page according to the features obtained. In the algorithm used by us, each feature contributes positively or negatively towards a page being classified to belong to a particular category. This feature is multiplied by the number of that type of features present in the document. The actual procedure is as follows:

Let W be a (cxn) weighed feature matrix, F be a (nx1) feature count matrix and V = WxF gives a cx1 matrix as shown below:

$$W = \begin{bmatrix} W_{00} & W_{01} & - & - & W_{0c-1} \\ W_{10} & W_{11} & - & - & W_{1c-1} \\ - & - & - & - & - \\ - & - & W_{ij} & - & - \\ W_{n-10} & W_{n-11} & - & - & W_{n-1c-1} \end{bmatrix} \quad F = \begin{bmatrix} F_{00} \\ F_{01} \\ - \\ F_{0i} \\ - \\ F_{0n-1} \end{bmatrix} \quad V = \begin{bmatrix} V_{00} \\ V_{01} \\ - \\ V_{0i} \\ - \\ V_{0c-1} \end{bmatrix}$$

In the matrix represented as W,

c - No. Of Categories

n – No. Of Features

And $W_{ij} \in [-1,+1]$ is the weight assigned to the i^{th} feature to contribute to the j^{th} category.

In the matrix F,

n - No. of features

F_{0i} is an integer count of i^{th} type of features present in the document.

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 is 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, we assigned different weights to each of these features for different categories. These assumptions are based on heuristics, which are later modified to reflect the likely values once the implementation runs on sample pages. For example, the presence of mathematical equations in an information page is very unlikely and hence it is assigned a negative weight for that category. Similarly, a high ratio of links is assigned a negative weight in the research page. 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. Pages containing less than 200 characters are excluded from being classified.

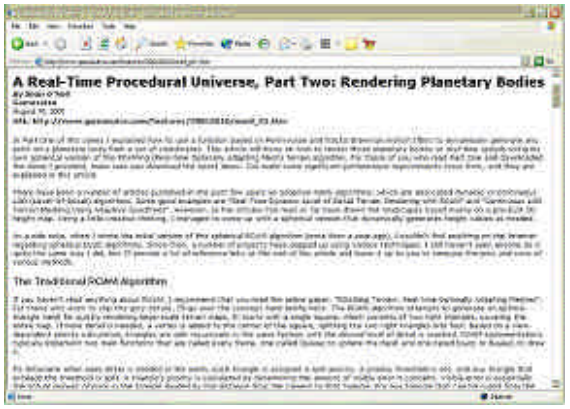

Figure1: wrongly categorized

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: Results of Implementation

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 might have fallen into information page. The reason can be attributed to the fact there is not a single link present in the page, 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 for effective categorization of web pages.

Our current implementation uses a method for classification, wherein the weights assigned to each feature are set manually during the initial training phase. A neural network based classification approach 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.

We have used images in addition to the structural information present in html 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 web 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] Oh-Woog Kwon, Jong-Hyoek Lee, Web page classification based on k-Nearest Neighbor approach

[3] 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.

[4] Andrew McCallum and Kamal Nigam, *A Comparision 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, *Hierarchial 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.*