Data Mining Techniques for Web Page Classification

Gabriel Fiol-Roig, Margaret Miró-Julià, Eduardo Herraiz

Math and Computer Science Department
Universitat de les Illes Balears
Ctra. de Valldemossa km. 7,5, 07122 Palma de Mallorca, SPAIN
{biel.fiol@uib.es, margaret.miro@uib.es}

Abstract Nowadays, the Web is an essential tool for most people. Internet provides millions of web pages for each and every search term. The Internet is a powerful medium for communication between computers and accessing online documents but it is not a tool for locating or organizing information. Tools like search engines assist users in locating information. The amount of daily searches on the web is broad and the task of getting interesting and required results quickly becomes very difficult. The use of an automatic web page classifier can simplify the process by assisting the search engine in getting relevant results. The web pages can present different and varied information depending on the characteristics of its content. The uncontrolled nature of web content presents additional challenges to web page classification as compared to traditional text classification, but the interconnected nature of hypertext also provides features that can assist the process. This paper analyses the feasibility of an automatic web page classifier, proposes several classifiers and studies their precision. In this sense, Data Mining techniques are of great importance and will be used to construct the classifiers.

Keywords: Data Mining, Artificial Intelligence, Decision Trees, Web page classification.

1 Introduction to the Problem

The growth of information sources available on the World Wide Web has made it necessary for users to utilize automated tools in finding the desired information resources. There is a necessity of creating intelligent systems for servers and clients that can effectively mine for knowledge. Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web by the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. There are roughly three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. Finally, web usage mining is the process of extracting interesting patterns in web access logs [1].

Web content mining is an automatic process that goes beyond keyword extraction. Since the content of a text document presents no machine readable semantic, some approaches have suggested restructuring the document content in a representation that could be exploited by machines. The usual approach to exploit known structure in documents is to use techniques to map documents to some data model. There are two web content mining strategies: those that directly mine the content of documents and those that improve on the content search of other tools. From a Data Mining point of view, Web mining, has three areas of interest: clustering (finding natural groupings of users, pages etc.), associations (which URLs tend to be requested together), and classification (characterization of documents).

Web page classification is the process of assigning a Web page to one or more predefined category labels [2]. Classification can be understood as a supervised learning problem in which a set of labeled data is used to train a classifier which can be applied to label future examples. The general problem of Web page classification can be divided into more specific problems. Subject classification is concerned about the subject or topic of a Web page. Functional classification cares about the role that the Web page plays. Sentiment classification focuses on the opinion that is presented in a Web page, that is, the author's attitude about some particular topic. Other types of classification include genre classification, search engine spam classification. This paper focuses on subject and functional classification.

The objective of this paper is to design an automatic classification system for web pages. The decision system designed has three main stages as indicated in Figure 1.1: the data processing phase, the data mining phase and the evaluation phase.

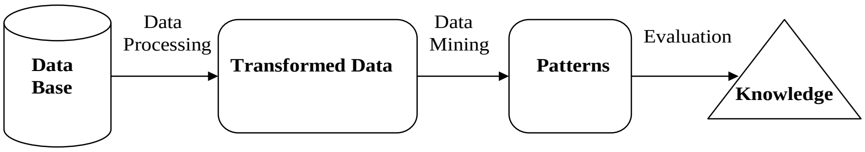


Fig. 1.1 The Decision System.

The data processing selects from the raw data base a data set that focuses on a subset of attributes or variables on which knowledge discovery is to be performed. It also removes outliers and redundant information, and uses HTML code to represent the processed data by means of an Object Attribute Table (OAT) [3]. The data mining phase converts the data contained in the OAT into useful patterns in particular decision trees are found [4]. The evaluation phase proves the consistency of pattern by means of a testing set. The positively evaluated decision system can then be used in real world situations that will allow for its validation.

2 The Data: Features of Web Pages

Web pages are what make up the World Wide Web. A Web page is a document or information resource written usually in HTML (hypertext markup language) and translated by your Web browser. Web pages are formed by a variety of information, such as: images, videos or other digital assets that are addressed by a common URL (Uniform Resource Locator).

These pages are typically written in scripting languages such as PHP, Perl, ASP, or JSP. The scripts in the pages run functions on the server that return things like the date and time, and database information. All the information is returned as HTML code, so when the page gets to your browser, all the browser has to do is translate the HTML, interpreting it and displaying it on the computer screen. Since all pages share the same language and elements it is possible to characterize each of them accurately in an automatic way.

All web pages are different, in order to classify them, data extracted from the HTML code will be used. Pages can then be classified according to multiple categories. Throughout this document, the following labels have been used: blog, video, images and news.

- Blog: short for weblog, is a personal online journal with reflections, comments provided by the writer. Blogs are frequently updated and intended for general public consumption. Blogs generally represent the personality of the author or reflect the purpose of the Web site that hosts the blog. Blogs can be distinguished by their structure: a series of entries posted to a single page in reverse-chronological order. Blogs contents are basically text, occasionally images and videos are included.
- Video: these web pages provide a venue for sharing videos among friends and family as well as a showcase for new and experienced videographers. Videos are streamed to users on the web site or via blogs and other Web sites. Specific codes for playing each video are embedded on the Web page.
- Image: a photo gallery on a website is collection of images or photos that is uploaded to a website and available for website visitors to view. These web pages hardly have any text and generally carry a navigator that allows the visitor to move around the images.

- News: an online newspaper is a web site which provides news on a basis which is close to real time. A good news site updates its content every few minutes.

The retrieval of the data available in web pages is not immediate and presents drawbacks. The extraction of useful data is a delicate process that requires careful handling. The use of an automatic program will simplify and improve the data acquisition process.

3 Transformed Data: Generation of the Object Attribute Table

The original data base is formed by a list of web pages. In the data processing phase an automatic program, written in Python, that processes the HTML code of the different web pages has been used. In particular, the program gathers information and generates the OAT. The attributes of the OAT, the columns, are the following:

- Page's text length (***TL***): number of text characters in the web page.
- External links (***EL***): number of links to external web pages.
- Internal links (***IL***): number of links to internal web pages.
- Image (***Im***): number of <img> elements in the web page.
- External Images (***EI***): number of external links with images.
- Internal Images (***II***): number of internal links with images.
- Multimedia Objects (***MO***): number of <object> elements in the web page, such as videos or flash player.
- Word Flash (***WF***): number of appearances of the word “flash” or similar in the text of the page.
- Word Video (***WV***): number of appearances of the word “video” or similar in the text of the page.
- Word Image (***WI***): number of appearances of the word “image” or similar in the text of the page.
- Word Blog (***WB***): number of appearances of the word “blog” or similar in the text of the page.
- Word News (***WN***): number of appearances of the word “news” or similar in the text of the page.

Table 3.1 illustrates the format of the OAT. Each row of the OAT describes the characteristics of a web page using the above defined attributes. There are a total of 12 numerical attributes and 271 rows. In the following step, an expert assigns classes to each of the rows according with the four categories mentioned above. The last column of Table 3.1 corresponds to the class.

Table 3.1 Object Attribute Table.

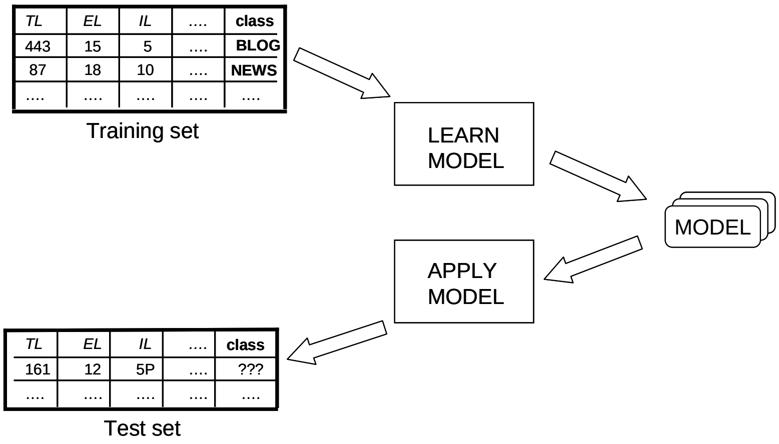
TL	*EL*	*IL*	*Im*	*EI*	*II*	*MO*	*WF*	*WV*	*WI*	*WB*	*WN*	*CLASS*
44299	96	38	20	20	0	0	0	3	166	2	2	IMAGE
54795	36	260	36	5	20	0	0	2	10	7	24	NEWS
235056	588	319	110	66	38	0	1	12	274	282	0	BLOG
84131	120	26	30	15	11	1	5	125	70	4	0	VIDEO
….	….	….	….	….	….	….	….	….	….	….	….	….

Some significant properties are connected to the use of OAT. Among these properties we single out the possibility to deal with binary or multivalued attributes [5], and to represent incomplete and vague knowledge [6, 7].

The OAT also facilitates the incorporation of an inferencial mechanism. This mechanism is based on the abstraction principle and considers the evolving characteristic of the environment. Moreover, it achieves a more efficient agent's decision stage.

4 Classification Methods

Classification methods allow the construction of a predictive model that is later used to assign labels to web pages. In order to create this model a training set must be used. The entries corresponding to 271 web pages are used to generate the model and evaluation is carried out using 20-fold cross validation. Figure 4.1 illustrates this process.



TL	EL	IL	….	class
443	15	5	….	BLOG
87	18	10	….	NEWS
….	….	….	….	….

Training set

TL	EL	IL	….	class
161	12	5P	….	???
….	….	….	….	….

Test set

Fig. 4.1 The Classification Process.

5 Decision Trees

Among available classification methods, decision trees were selected for their simplicity and intuitiveness. Decision trees were developed using WEKA [8], (Waikato Environment for Knowledge Analysis) a collection of machine learning algorithms for data mining tasks.

Different classification algorithms were tried out, among them Best First Decision Tree (BFT), Logistic Model Tree (LMT), J48 Graft Tree (J48GT) and J48 Pruned Tree (J48PT). Obviously, the trees obtained by the different methods differ. Essentially, this difference becomes apparent in the tree's complexity and its precision.

The classification methods were applied to the set of 271 instances and 12 variables. The results of some of the most relevant trees, using 20-fold cross validation, are shown in Table 5.1.

Table 5.1 Classification Results.

METHOD	Correct	Incorrect	Correct (%)	Incorrect (%)
BFT	246	25	90.78	9.22
LMT	245	26	90.41	9.59
J48GT	249	22	91.88	8.12
J48PT	245	26	90.41	9.59

The confusion matrices, shown in Table 5.2, indicate the number of classified instances according to class. The notation used for the classes is the following **a** = IMAGE, **b** = NEWS, **c** = BLOG and **d** = VIDEO.

Table 5.2 Confusion Matrices.

	BFT					**LMT**			
	a	**b**	**c**	**d**		**a**	**b**	**c**	**d**
a	61	2	3		**a**	60	2	2	2
b	4	83	3	0	**b**	5	83	2	0
c	3	2	78	1	**c**	3	0	77	4
d	3	2	2	24	**d**	2	1	3	25

	J48GT					**J48PT**			
	a	**b**	**c**	**d**		**a**	**b**	**c**	**d**
a	60	3	2	1	**a**	61	3	2	0
b	1	86	3	0	**b**	5	82	3	0
c	3	1	79	1	**c**	4	1	78	1
d	3	2	2	24	**d**	3	2	2	24

The results of some of the most relevant trees are shown below in Figure 4.1 and Figure 4.2.

The LMT tree has a total of 11 nodes, 5 internal nodes corresponding to the 4 attributes used and 6 leaf nodes or branches representing the classes. The average branch length is 3 and the maximum depth is 4. The attributes representing the internal nodes are: ***WN***, ***EL, WV*** and ***TL***.

The J48 pruned tree has a total of 19 nodes, 9 internal nodes corresponding to the 6 attributes used and 10 leaf nodes or branches representing the classes. The average branch length is 3.5 and the maximum depth is 5. The attributes representing the internal nodes are: ***WN***, ***EL, TL, WV, Im*** and ***WI***.

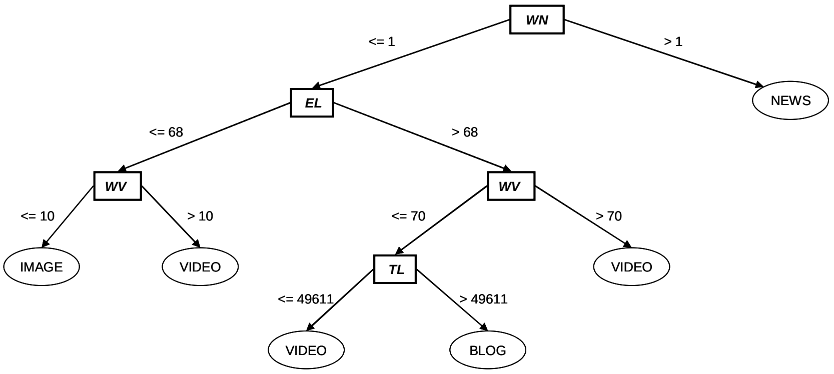


Fig. 4.1 The Logistic Model Tree.

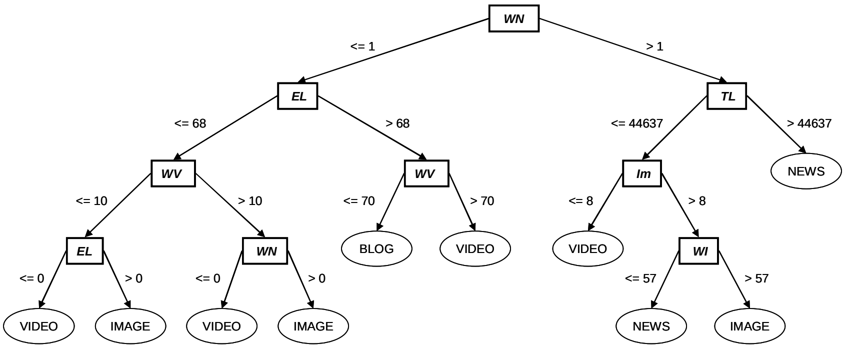


Fig. 4.2 The J48 pruned Tree.

If both trees are compared, it can be seen that attributes ***WN*** and ***EL*** appear at the upper levels of the LMT tree and the J48 pruned tree suggesting their importance. The number of appearances of the word “news” (***WN***) is a deciding factor for the classification of web pages. None of the pages with 1 or less appearances of the word “news” is a news page. On the other hand, a high number of external links and a reduced number of appearances of the word “video” correspond with a blog page.

6 Conclusions and Future Work

Classification methods based on data mining techniques have been applied to the area of web mining in order to construct efficient trees that classify web pages depending on their features.

In the development of the method, several steps must be considered. First of all, raw data must be collected and analyzed. The selection of data is an arduous and necessary task and requires an in depth analysis of all the problem requirements. Secondly, a variable analysis process takes place. Those unnecessary variables for data classification are removed and, if necessary, new variables are considered. Once the data is consolidated, the data’s final format in terms of an Object Attribute Table (OAT) is introduced. Third, data mining methods are applied to the OAT, in order to find valid and efficient models that classify the instances. The resulting trees are analyzed in terms of complexity and precision. Generally speaking, classification accuracy requires complex trees.

The results show that a web page classification based on web page features available in the page’s HTML code is possible. The success rate is acceptable even though this paper offers a limited vision of one of the many solutions available. The following aspects can be considered as future work: other different variables, such as page style or text style may offer better results and should be studied. Also other decision trees can be obtained using different learning algorithms and should be evaluated.

References

[1] Dunham M. H. (2003) Data Mining. Introductory and Advanced Topics, Prentice Hall, New Jersey.

[2] Qi X and Davison B.D. (2009) Web Page Classification: Features and Algorithms. ACM Computing Surveys, Vol. 41, No. 2, Article 12.

[3] Miró-Julià M., Fiol-Riog G and Vaquer-Ferrer D (2009) Classification using Intelligent Approaches: an Example in Social Assistance. Frontiers in Artificial Intelligence and Applications 202: 138-146.

[4] Fiol-Roig G (1999) UIB-IK: A Computer System for Decision Trees Induction. LNCS 1609: 601-611.

[5] Miró-Julià M and Fiol-Roig G (2003) An Algebra for the Treatment of Multivalued Information Systems. LNCS, 2652: 556-563.

[6] Fiol-Roig G (2004) Learning from Incompletely Specified Object Attribute Tables with Continuous Attributes. Frontiers in Artificial Intelligence and Applications 113: 145-152.

[7] Miró-Julià M (2005) Degenerate Arrays: a Framework using Uncertain Data Tables. LNCS 3643: 21-26.

[8] Ian H. Witten and Eibe Frank Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.