Detecting Blogs Independently From The Language And Content

Francisco Manuel Rangel Pardo[1], Anselmo Peñas Padilla[2]

[1] Dpto. I+D+i, Corex Soluciones Informáticas, SL
Valencia - Spain
francisco.rangel@corex.es
[2] Dpto. Lenguajes y Sistemas Informáticos, UNED
Madrid - Spain
anselmo@lsi.uned.es

Abstract: This paper obtains a high performance representation in the automatic classification of Blog pages, regardless of its content, style, author and language. The research is based on the concept of "frame" to represent the knowledge that makes a page of Blog type recognisable to everybody at first sight. We have built a dataset in four different languages with varied contents in fifteen different categories, we have experimented with four different representations and evaluated with four different methods of inductive learning. The result of the proposed method is a high performance, with F values higher than 0.950 and an interval of error lower than 2%. This shows the superiority of this method compared with others and its independency from both language and content in which the page is written.

Keywords: Blogs, Classification, Web Classification, Web Clustering, Blog Search, language independent classification, multilingual classification

1 Introduction

We have focused the investigation on simulating the cognitive process of visual recognition of a Blog to obtain a representation able to determine, through an inductive learning method, if a page is or is not of this category.

The revolution of Internet toward a collaborative model, where everyone can contribute knowledge, experiences and opinions freely and globally, has raised in the research a growing interest in everything related to social media mining, opinion mining and sentiment analysis.

It is therefore vital to indentify the sources to obtain that information and opinions, and one of the most important sources for that task are blogs, because of its freedom for expression, its ease for collaboration and its rapid growth in Web 2.0.

But Internet is global, what means its contents are expressed in a high quantity of different languages, so it is really important to find a method rich enough in representation, in order to obtain an efficient classification of this kind of pages, as we have said before, independently from its language or content.

The main contribution of this paper is that we have focused on obtaining the visual characteristics that make them recognizable to a first sight for a human, obtaining a

novel representation model that obtains high values of statistical-F and low levels of error, significantly improving the results front the state of the art methods.

This paper has been structured in a review of related work in section 2. Under section 3 we present our proposal based on the visual characteristics of Blog pages and the way to obtain features to the page representation. In section 4 it has been described the creation of test collection and its characteristics. Then in section 5, we describe the methodology and the theoretical framework of experimentation and the evaluation of the results. In section 6 we discuss the experimental results obtained, with the intention to make way for the conclusions and the future work done in section 7. Thereupon, we attach the references to the wording of the paper.

2 Related Work

The state of the art shows that a large number of alternatives has been explored in the extraction of features of the Web for its future treatment, learning and modelling.

We have investigated in different formal representations of the pages: from the classical model of Bag Of Words (BoW), based on obtaining features from the content of the page; to models exploiting "links structure" and "meta-data" from the Web, and the relationship between pages.

Studies like [16] show us how the classification of pages based on summaries made by humans, have a significant enhancement, reducing the dimensionality and delimiting the problem in a more concrete vocabulary, achieving improvements of 12,9% over the baseline BoW.

[12] proposes a classification based only in structural features, achieving a 92% in statistical-F in classification of first level domains.

[11] makes a classification from the words that appear on the URL of the page, obtaining very interesting results, in a similar way than that made by [17] in order to block "spot advertisings" depending on its URL.

Other lines of investigation are an approach to the contextual analysis of the pages, which are divided on three main lines: hypertext analysis, links analysis and neighbourhood analysis.

Hypertext analysis is based on drawing characteristics from the anchor text, the headings, the pointed out pages, and generally all those elements which were highlighted in some way in the html structure. Thereby [19] uses a combination of characteristics drawn from the title of the pages and from the anchorages, showing an increase of performance facing the classical methods based on content.

The approach based on links analysis, combines the analysis described above with the textual analysis of the referenced pages. [3] achieves an increase of 46 points in the F1 on the textual analysis of the pages. [18] makes use of the HITS algorithm to explore the topology of hyperlinks, and [9] who uses the combination of kernel functions for support vector machines to utilize jointly the textual information and the cocitation analysis.

[4] and [14] make use of the neighbourhood analysis, it means, they use the classification of the neighbour pages to determine the category of the new documents.

Regarding the classification of Blog-type pages, we have not found a high number of papers despite an incessant focus of interest in Opinion Mining, Sentiment Analysis and other kind of Mining Social Media.

In [15] find a use of the meta-information and the links on the page in order to classify it into a specific domain where content is more semantically similar than into a first level domain, and they achieve better results in the classification of blogs than the compared methods.

The NITLE project (http://www.knowledgesearch.org/census/index.html) to census the existing blogs in the Web, uses several heuristic rules like that the word "blog" appears more than five times, that the page has been created by an automatic generator of content, that its URL belongs to a known publisher of blogs (blogger or wordpress), that the page has got or allows subscription via RSS or atom, and several more, but it does not provide information on the obtained results.

In trade there are many blog search engines like Technorati, Agregax or Google. All of them follow the philosophy of indexing and publish big sites of content generation like Blogger, or publishing all pages that have got rss or atom subscription. For more information you can visit (http://www.google.com/intl/es/help/about_blogsearch.html#whichblogs).

3 Visual characteristics of the Blogs

Perhaps, Blogs are one of the most heterogeneous sorts of pages in relation to its content that might exist in the Web, due to the fact that they can deal with a high variety of themes, even inside the same Blog. So a priori it would be difficult to obtain a good representation based on the content, besides of the problems emerging with that type of representations when it deals with multilingualism.

In the same way, the use of fixed rules relating to the underlying technology of creation of Blogs (e.g. Meta-generator or containing subscription); or the use of logical rules of engagement (e.g. specific urls of generators of Blogs or the word "blog" appears more than n times); does not seem to be appropriate because on one hand they depend extremely on the underlying technology, limiting the possibility to adapt itself to new situations, and on the other hand they confuse the concept of Blog as collaborative page and not as a page which publishes contents that could be subscribed, which it can also be a newsgroup, CMS, forums and wikis, among others.

In the other way it is undeniable that somebody who has seen a series of Blogs, would be able to identify, in the vast majority of cases, a new Blog as such, independently from the content, author, style or language in which it has been written.

This spontaneous identification answers to a cognitive process of vision (visual cognitive process) and a recognition of a pattern which can be described through the concept of "frame" introduced by Marvin Minsky [13]

In this paper we have made a "conceptual" use of "frame" as a representative method (or container) of knowledge, it means, as a method of representing the pages. Then, we use a supervised inductive learning model to assign values to the frame slots (features of the learning model), so that once built the model may be able to predict the class of a given page as an example and represented in this way.

The visual characteristics would be the features that a human would be able to indentify at a glance:

- Blocks of information, the posts, structured as entries of a diary, with a date of publication, a headline, a content and the possibility to include comments to allow some feedback the conversation and collaboration.

- A blogroll such a group of links to the same website and with the aim of providing direct and permanent access to past entries.
- Words highly representative because their high frequency of appearance, such as blog, post, rss, atom and comment.

The supervised inductive learning will allow a flexible identification of blogs that have any of the above features, for example, to identify a blog without rss subscription such as it is a private blog, or to reject a page that would contain it but actually it is a newsgroup.

We have identified entities such as dates or comments by using regular expressions specifically built for this task. To ensure the independence of the language (only for the languages we are testing), we have created regular expressions to identify the different ways that an entity can appear in each language, for example, in Spanish we could have the next forms "21/12/2009" and "21 de diciembre de 2009", but in English would be "12/21/2009".

We implement the three visual characteristics described with a set of 14 features based on ratios of occurrence of some entities over others. We do not save boolean values, we save ratios and the learning method determines the margins of the combined features to determine the class. The obtained features are the following:

- Frecuency of occurence of the word blog in the Url
- Frecuency of occurence of the word blog in the document
- Frecuency of occurence of the word post in the document
- Frecuency of occurence of the word RSS or ATOM in the document
- Ratio between number of comments and number of dates, obtained by regular expressions named above
- Ratio between number of comments into a link and number of dates
- Ratio between number of comments into a link and total number of comments
- Ratio between number of comments and number of headlines. The number of headlines is obtained with a regular expression that extracts entities between the different tags <H>
- Ratio between number of coments into links and number of headlines
- Ratio between number of dates and number of headlines
- Boolean indicating if there is a blogroll. We search for HTML containers such as <UL> or <OL>
- Ratio between number of links to the same domain and number of links to different domain, located them in the blogroll
- Ratio between number of links to different domain located in the blogroll and total number of links in the page
- Ratio between total number of links in blogroll and total number of links in page.

4 Training Set

We have built a training set with sites classified as Blog and No-Blog, in four different languages: English, Spanish, French and German. We have obtained this collection from the sites listed manually on the directory DMOZ ODP.

We have obtained the Blog pages from concrete categories listed as Blogs under each subdirectory of languages from DMOZ, searching them manually and executing a crawl from the main category Blog, obtaining the quantity shown in Table 1:

LANGUAGE	Nº PAGES
English	1859
Spanish	785
German	508
French	668

Table 1: Number of blogs by language

No-Blog pages have been obtained by making a crawl of the indexed pages in the DMOZ, under each language and in each of the categories of the subdirectory (e.g.: Arts, Bussiness, Computers and several more) and avoiding the download of pages of the subcategory Blog, and including personal and/or corporative pages as well as groups of news, forums or wikis.

We have preprocessed manually and automaticaly the collection, deleting pages without relevant information, such as removed pages, error pages, non-existing pages, redirected pages and any kind of error pages.

We have made a balancing of the total of pages of Blog type and No-Blog type depending on the language. The total of Blog pages is inferior, so we have performed a random selection of pages No-Blog to obtain a similar quantity of totals as Blog pages, everything depending on the language, and maintaining a similar quantity of pages for each subcategory (Arts, Business or Compuers, for example) as it is shown in the table below (2):

LANG.	PAG.	CATEG.	PAG/CAT
English	1941	13	150
Spanish	774	13	60
German	554	14	40
French	674	15	45

Table 2: Number of Not-Blog by language, categories and pages per category

The final collection is as follows:

Lang./Class	Blog	Non-Blog	TOTAL
English	1859	1941	3800
Spanish	785	774	1559
German	508	554	1062
French	668	674	1342
TOTAL	3820	3942	7763

Table 3: Final Test Collection

We have created a collection for each language and a collection with all languages in common, to experiment with each language and testing multilingual.

You can obtain this collections on http://www.wikimasd.com.

5 Methodology and evaluation framework

5.1 Election of the classifier

The methods of inductive learning allow the construction of models that generalize the behaviour of the data given as evidence to predict new data. In the case of this paper, these methods are binary classifiers able to distinguish between the pages belonging to the category of Blog or not.

To test whether the results are independent from the learning method used, we have experienced with four different learning methods, Naïve Bayes, BayesNet, Support Vector Machines and decision trees, in their respective implementations on Weka.

5.2 Technique and measures of evaluation

In the case of classifiers, there are diverse techniques to learn, from which we choose the evaluation of hypothesis based on accuracy, where the percentage of error made between the formulated hypothesis and the real value is assessed, and the learning is guided to minimize the number of committed errors [8]. Finally, we make an evaluation of the results through Cross Validation.

The evaluation done through Weka is based on accuracy, obtaining statistical-F through the calculation of the harmonius media from Precission and Recall.

Thereupon, we have made the t-student test from two tails of the statistical F series of the methods (of learning and of representation), comparing them to a level of signification of 95%. The Null Hypothesis H_0 will be that the pair of compared series is equal and for that reason the performance of the compared representations will be equal. Nonetheless if the called statistic is superior to a tabulated value, the Null hypothesis will be rejected, so one of the methods of representation is superior to the other. Depending on the sign we will conclude which of them.

5.3 Measures of error

To establish a confidence level of evaluation allows us, given a sample named S with n instances taken from an objective function f with a distribution D, establish some confidence intervals for the real error (error(h)) from a hypothesis of sample error (errorS(h)). Consequently, to a level of confidence c%, we can determine the interval error as:

$$errorR(h) = errorS(h) \pm z_c \sqrt{\frac{errorS(h)(1-errorS(h))}{n}} \quad (1)$$

Where z_c is obtained through the normal distribution and the used value in the different evaluations is of 1.96 equivalent to 95% of certainty.

5.4 Baseline and other representations

We can use the majority baseline formed from the ratio between "blog or no-blog / total" such as base to compare all other baselines and our method of representation.

We have token as baseline representation that made by Google Blog-Search, consisting in the determination of if a page is or not of Blog type having subscription (RSS or Atom).

We have obtained a representation based on the standard "bag of words BoW", using as a characteristic, the most representative words of its corpus, using a previous process of stem of Porter and filtering through stopwords and selecting the 10% of apparition more frequent.

We have obtained the representation of the project NITLE, replacing its fixed rules by an inductive learning of the corresponding characteristics to those rules (e.g. Contains subscription, it has been automatically generated, among others explained on section 2)

Finally we have called our proposed method of representation, CRX.

The number of characteristics of each of the representations, by language, is as follows:

	EN	ES	DE	FR	MULT
BoW	686	1665	661	798	2484
GOOGLE	1				
NITLE	6				
CRX	14				

Table 4: Dimensionality of the representations

Spanish BoW representation has more features than other languages because we are chosen in all cases those words that had a 10% of frequency of occurrence, and Spanish language has had more words that carry out that condition.

6 Experimental results

We have conducted several experiments consisting on the comparison of the obtained results through the different learning methods for the different representations and in each one of the languages, obtaining the statistical F and comparing the series of values F through a t-student test.

In the same way, we have experimented with the whole collection, obtaining the number of committed errors per each method and the interval of real error in a confidence level of 95%.

The hypothesis H_0 is that the proposed representation does not get a significant increase in the performance of classifiers. The fact of rejecting the hypothesis means a significant increase of performance, regardless to the language, content and method of training.

We will validate H_0 (accepting or rejecting it) through the T-student test.

In the table shown below (Table 5), we can see the results for the values of F obtained in each classification for each representation. The values in each cell correspond to the obtained F, on the left we see the case of positive classification as Blog,

and on the right we see the classification as No-Blog. We mark in bold the best results obtained into a same classification for each language.

		NAÏV	BNet	SVM	J48
English	Majority	0.489 / 0.511			
	BoW	0.762 / 0.774	0.847 / 0.852	0.872 / 0.878	0.940 / 0.935
	Google	0.874 / 0.881			
	Nitle	0.926 / 0.932	**0.926** / 0.918	0.905 / 0.918	0.935 / 0.931
	Crx	**0.970 / 0.971**	0.924 / **0.932**	**0.952 / 0.957**	**0.980 / 0.980**
Spanish	Majority	0.503 / 0.497			
	BoW	0.834 / 0.807	0.824 / 0.815	0.918 / 0.906	0.939 / 0.921
	Google	0.893 / 0.891			
	Nitle	0.928 / 0.931	0.929 / 0.926	0.941 / 0.939	0.942 / 0.942
	Crx	**0.989 / 0.989**	**0.941 / 0.944**	**0.973 / 0.973**	**0.993 / 0.993**
German	Majority	0.478 / 0.522			
	BoW	0.835 / 0.818	0.904 / 0.890	0.866 / 0.866	0.941 / 0.932
	Google	0.891 / 0.901			
	Nitle	0.923 / 0.931	0.923 / 0.931	0.907 / 0.925	0.925 / 0.934
	Crx	**0.963 / 0.970**	**0.927 / 0.944**	**0.945 / 0.959**	**0.961 / 0.970**
French	Majority	0.498 / 0.502			
	BoW	0.820 / 0.759	0.850 / 0.835	0.886 / 0.874	0.927 / 0.907
	Google	0.907 / 0.908			
	Nitle	0.951 / 0.952	0.952 / 0.953	0.956 / 0.959	0.952 / 0.955
	Crx	**0.986 / 0.986**	**0.953 / 0.956**	**0.973 / 0.975**	**0.986 / 0.987**
Multi.	Majority	0.492 / 0.508			
	BoW	0.686 / 0.741	0.701 / 0.755	0.876 / 0.875	0.941 / 0.931
	Google	0.886 / 0.891			
	Nitle	0.936 / 0.939	0.922 / 0.928	0.922 / 0.931	0.938 / 0.941
	Crx	**0.980 / 0.981**	**0.935 / 0.942**	**0.960 / 0.964**	**0.984 / 0.985**

Table 5: Statistical-F validation for each language and each representation

Table 5 shows superior results in almost all cases (except only one case) for the proposed representation. Furthermore, in multilingual case the proposed representation maintains similar values for statistical-F , meanwhile the rest, especially BoW, declines slightly its performance.

There is a pattern with Naïve Bayes and BayesNet, the results for all representations and languages are lower compared with decision trees or support vector ma-

chines. Furthermore, the difference between our method and the best method compared when BayesNet is used, is a little more than with other methods.

To determine that there is a statically significant improvement we use the t-student test (Table 6) with the series of values of the proposed method compared with NITLE, the method with best results obtained compared to the rest (because its average is higher than the others):

Language	NITLE vs. CRX
English	t = 5,29 > 2,365
Spanish	t = 6,07 > 2,365
German	t = 6,23 > 2,365
French	t = 4,27 > 2,365
Multilingual	**t = 7,20 > 2,365**

Table 6: t-student test paired with two tails for the F series for each languages for NITLE and CRX

By comparing the t-student test with the NITLE method, the best classifier in the group, the hypothesis H_0 is rejected for all cases, what means that the proposed method obtains significantly better results than others compared and it demonstrates the superiority of it.

As it is shown in table 6, the higher value of the t-student, therefore the most significant difference between representations, is done with multilingual training, what indicates that proposed representation is much better than others dealing with multilingual data.

The results of Google for the same experiment (language) are always the same, due to their dependence to the apparition of a single rule: “if the page has subscription, it is a blog”, similarly with majority baseline.

The results of NITLE with the learning methods are higher than those expected to be from its version with rules, but this cannot be corroborated, due to the capacity of handling inconsistency and uncertainty that learning methods incorporate.

Dimensionality in the case of BoW, does practically unmanageable the training (see Table 4).

The above results (Tables 5 and 6) show (the value obtained in all cases for the t-student test is higher than the tabulated value) that all the classifiers for all languages have a significantly higher yield in the case of the proposed representation.

A comparative analysis of the interval error, only for the case of multilingual training and the Naïve Bayes learning method, shows that the number of mistakes and the intervals obtained for the proposed representation are significantly lower than those obtained by other methods, and it gives margins of error less than 2%:

	Nº errors	Interval
Majority	Not calculated	
BoW	2205	0,2840+-0,0109
Google	862	0,1115+-0,0070
NITLE	487	0,0627+-0,0054
CRX	149	0,0192+-0,0031

Table 7: Interval error in multilingual validation

Because space reasons we only have reflected the results obtained for multilingualism, on the other hand, the main objective of the paper. The results show the large decline in the number of errors from approximately 30% of BoW representation to less than 2% of our proposed method, passing through Google with 11% and Nitle with 6%.

The BoW results are due to the high dimensionality, multilingualism and variety of content, which make difficult the learning task, so it is known that these methods are not suitable for this kind of problems.

The results of Google and Nitle are more interesting to analyse. Google works with a fixed rule that says "if a page has got subscription is a blog", which causes many failures with pages that have subscription but are not blogs, such as wikis, forums or newsgroups, and many blogs that do not have subscription. Nitle reduces such errors by adding rules that define the conditions for the page to be a blog, but it often fails in cases where the rules are met but it is not a blog (e.g. a personal web page generated with blogger) and in cases where the rules are not met but it is a blog (e.g. a blog created programmatically and hosted on an own domain).

7 Conclusions and future work

This work has been framed on the lines of current research to propose a new way to characterize or represent web pages based in the cognitive process of visual recognition to classify them as a Blog or not-Blog.

We have created a test collection in four different languages with contents in different domains and annotated in two discriminating classes, Blog and no-Blog.

We have shown that the proposed representation, obtains a significantly higher value on statistical-F than four compared methods of the state of the art. Results are higher than 0,920 points in F in all cases, and we have reduced the interval of error to lower levels than 2%.

We have found that word-based methods are not suitable mainly for two reasons, firstly because they rise rapidly its dimensionality in learning, and secondly, and partly due to what has been said in the first reason, because they reduce their performance.

If we compared the proposed method with that of Google, our method distinguishes between pages of type Blog and any other type of page that would have subscription RSS like news, forums or wikis.

Comparing the proposed method with NITLE, our method is based on inductive learning instead of logic rules, which allows deal with uncertain and incomplete information and adapting itself to any change it could suffer, without the necessity of retraining.

Our method is not based on the underlying technology but on the concept of Blog, which allows the method to break away from the technology and adapt to changes in the same without the need for retraining.

The proposed method prioritizes Blogs containing reviews instead of those which not have. This allows a better identification of those which are truly collaborative and that have a major interest to researches on Opinion Mining and Sentiment Analysis.

The main conclusion of the paper is that we have obtained a representation conforming to the "frame concept" for the identification of a blog by a human, which it

makes possible to identify them independently of their content, style, author and language.

Future work would be to extend the representation to deal with new languages. To achieve this purpose we would have to strengthen the method to obtain entities such as dates or comments, improving the syntax of the regular expressions which obtain characteristics or using new methods to entity recognition. It would be interesting to include temporal analysis of the posts or adding some logical rules such as NITLE Project.

Acknowledgements

This paper has been partially subsidized by QEAVis-Catiex(TIN2007-67591-C02-01) project of the Ministry of Science and Innovation and partially by R&D department from Corex Soluciones Informáticas.

References

1. Attardi, Giuseppe; Gulli, Antonio; Sebastiani, Fabrizio. Automatic Web Page Categorization by Link and Context Analysis. Proceedings of THAI-99, European Symposium on Telematics, Hypermedia and Artificial Intelligence, pp. 105-119, 1999
2. Barzilay, R., Elhadad N., McKeonwn K. R. Inferring Strategies for Sentence Ordering in MultiDocument News Summarization. Journal of Artificial Intelligence Research, pp. 17:35-55, 2002
3. Calado, Pável; Cristo, Marco; Moura, Edleno; Ziviani, Nivio; Ribeiro-Neto, Berthier; Adré Goncalves, Marcos. Combining Link-based and Content-based Methods for Web Document Classification. Proceedings of the twelfth international conference on Information and knowledge management, pp. 394 – 401, 2003
4. Chakrabarti S, Dom B, Indyk P. Enhaced Hypertext Categorization Using Hyperlinks. In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 307-318, 1998
5. Cristo, Marco; Calado, Pável; Silva de Moura, Edleno; Ziviani, Nivio; Berthier, Ribeiro-Neto. Link Information as a Similarity Measure in Web Classification. Journal of the American Society for Information Science and Technology, pp. 208 – 221, 2006
6. Furnkranz J. Exploiting Structural Information for Text Classification on the WWW. Intelligent Data Analysis, pp. 487-498, 1999
7. Glover E.J, Tsioutsiouliklis K., Lawrence S, Pennock, D.M., Flake G.W.. Using Web Structure for Classifying and Describing Web Pages. Proceedings of WWW-02. International Conference on the World Wide Web, pp. 562 – 569, 2002
8. Hernández Orallo, José; Ramírez Quintana, Mª José, Ferri Ramírez, César. Introducción a la Minería de Datos. 2005. ISBN: 84-205-4091-9
9. Joachinms T., Cristianini N., Shawe-Taylor J. Composite kernels for hypertext categorisation. In C. Broodley and A.Daniluk, editors, Proceedings of ICML-01, 18th International Conference on Machine Learning, pp. 250-257, 2001
10. Joachims. Learning to Classify Text Using Support Vector Machines. Methods, Theory and Algorithms. 2002. ISBN 0-7923-7679-X
11. Kan, Min-Yen. Web Page Classification Without the Web Page. Proceedings of the 13th International. World Wide Web Conference Alternate Track Papers & Posters, pp. 262-263, 2004
12. Lindemann, Christoph; Littig, Lars. Classifying Web Sites. Proceedings of the 16th international conference on World Wide Web, pp. 1143 – 1144, 2007

13. Minsky, Marvin. Frame-system theory Thinking. Readings in cognitive science, 1977
14. Oh H.J, Myaeng S.H., Lee M.H. A practical Hypertext Categorization Method Using Links and Incrementally Available Class Information. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 264-271, 2000
15. Rangel Pardo, Francisco Manuel; Peñas Padilla, Anselmo. Clasificación de páginas Web en Dominio Específico. Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN, pp. 89-97, 2008
16. Shen, Dou; Chen, Zheng; Yang, Qiang; Zeng, Hua-Jun; Zhang, Benyu; Lu, Yuchang; Ma, Wei-Ting. Web Page Classification Through Summarization. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 242 – 249, 2004
17. Shih, L.K; Karger, D.R. Using Urls and Table Layout for Web Classification Tasks. Proceedings of the 13th international conference on World Wide Web, pp. 193-202, 2004
18. Slattery S., Craven M. Discovering Test Set Regularities in Relational Domains. Proceedings of ICML-00, 17th International Conference on Machine Learning, pp. 895-902, 2000
19. Sun, Aixin; Lim Ee-Peng; Ng, Wee-Keong. Web Classification Using Support Vector Machine. Proceedings of the 4th international workshop on Web information and data management, pp. 96-99, 2002