Automatic Categorisation of Croatian Web Sites

Jasminka Dobša, Danijel Radošević, Zlatko Stapić, Marinko Zubac
Faculty of organization and informatics
University of Zagreb
Pavlinska 2, Varaždin, Croatia
Tel.: +385 42 213 777, Fax: +385 42 213 413, E-mail: *jdobsa@foi.hr*

Abstract - On the Web site www.hr we can find the catalogue of Croatian Web sites organized hierarchically in more then 600 categories. So far new Web sites have been added into the hierarchy manually. The aim of our work was to research the possibilities of automatic categorisation of the Croatian Web sites in the hierarchy of catalogue. For the representation of documents (Web sites) we have used text mining technique of bag of words representation, while for the purpose of categorisation we have used the technique of support vector machines. The experiments are conducted for categorisation of Web sites in 14 categories on the highest hierarchical level.

I. INTRODUCTION

On the Web site www.hr we can find the catalogue of Croatian Web sites organized hierarchically in more then 600 categories. So far new Web sites have been added into the hierarchy manually. The aim of our work was to research the possibilities of automatic categorisation of the Croatian Web sites in the hierarchy of catalogue. That would enable a much more comprehensive recall of Croatian Web sites in the hierarchy. The categories are organized hierarchically in a few levels and categories on the higher hierarchical levels contain subcategories and pseudosubcategories on the lower hierarchical level. Pseudosubcategory of a certain category C is a category which is originally contained in some category C^*, but also associated to category C. Our experiments are conducted for categorisation of Web sites in 14 categories on the highest hierarchical level.

Hierarchy of Croatian Web pages is an example of Yahoo!-like hierachical catalogues. Document indexing and categorisation of such a catalogues was discussed in [1] , [3], [7], [10], and [11] . Web pages are special kinds of documents, as they consist not only of a text, but also of a set of incoming and outgoing pointers. Attardi et al. [1] propose an indexing technique specific to Web documents which is based on the notion of the **blurb** of a document. Given a test document d_j, $blurb(d_j)$ is a document which is formed of the juxtaposition of the text windows containing hypertextual pointers from other documents to document d_j. Fuhr et al. [5] propose representation of document d_j by indexing of document d_j itself and the documents directly pointed by d_j. This choice is motivated by the fact that usually we are interested in categorization of Web sites, rather then in categorization of Web pages. Root page of a Web site is often content less, because it is just a collection of entry points to children pages. The experiments in this paper are based on this approach.

For the representation of documents (Web sites) we have used text mining technique of **bag of words representation**. This technique is implemented by a **term-document matrix**, which is constructed on the basis of frequencies of index terms in documents. The bag of words representation is described in the second section. For the experiment we have used sites already present in the hierarchy. The process of classification, evaluation, and classification algorithm of support vector machines are described in the third section. In the forth section we give the results of our experiment. We have compared performance of the classification for two different representations of Web sites. The first representation is formed by xml file associated to the www.hr site using the description of every site provided in the hierarchy, and the second representation is formed by extended xml file where sites are represented by the text present on the first level of links contained on the site. In the sixth section we make conclusions and a discussion.

II. BAG OF WORDS REPRESENTATION

The bag of words representation or the vector space model [2] is nowadays the most popular model of text representation among the research community of text mining. The name "bag of words" is due to the fact that in this representation the documents are represented as a set of terms contained in the document, which means that dependence between the terms are neglected and it is assumed that index terms are mutually independent. This seems as a disadvantage, but in practice it is shown that indiscriminate application of term dependence to all the documents in the collection might in fact hurt the overall performance [2]. The technique of bag of words representation is implemented by a term-document matrix. First we have to create the list of index terms. This list is created from terms contained in the documents of collection by rejecting the terms contained in less then n_1 and more then n_2 documents and terms contained in the stop list of terms for certain language. Stop terms are terms which appear very often in the text (like articles, prepositions and conjuctions) and the appearance of which does not have any discriminate meaning. The term-document matrix is $m \times n$ matrix $A = [a_{ij}]$ where m is number of terms, n is number of documents, and a_{ij} is weight of i-th term in the j-th document. The weight of the term in the document has a local and a global component or **term frequency (tf)** and **inverse document frequency (idf)** factor [13]. Term frequechy factor depends only on the frequency of the term in the document, while inverse term frequency depends on the number of documents in which the term appears. This model of term weighting in

the documents assumes that, for a certain document, terms with the best discrimination power are terms with high term frequences in the document and low overall collection frequences. For the purpose of our experiment we have used popular TF-IDF weighting formula in which term frequency is simply the frequency of the term in the document and inverse document frequency is calculated by formula $\log(N/n)$, where n is the number of documents in which term is contained and N is the number of documents in the collection. Further, term weights a corrected by normalizing columns of the term-document matrix to the unit norm. The columns of the term-document matrix represent documents and their normalization neutralize the effect of a different length of documents.

III. CLASSIFICATION AND ITS EVALUATION

A. A Definition of Text Categorization

Text categorisation [14] is the task of assigning a Boolean value to each pair $(d_j, c_i) \in D \times C$, where D is the domain of documents and $C = \{c_1, ..., c_k\}$ is the set of predefined categories. If document d_j belongs to category c_i, then we will assign 1 to the pair (d_j, c_i), otherwise we will assign 0 to that pair. More formally, the task is to approximate the unknown target function $\widetilde{\Phi} : D \times C \to \{0, 1\}$, that describes how documents ought to be classified, by means of a function $\Phi : D \times C \to \{0, 1\}$ called the **classifier** (or **rule**, **hypothesis**, **model**) such that $\widetilde{\Phi}$ and Φ coincide as much as possible. The case in which exactly one category must be assigned to each $d_j \in D$ is called the single-label case, while the case where one ore more categories may be assigned to the same $d_j \in D$ is called the multi-label case. A special case of a single-label case of text categorization is binary text categorization, in which each document must be assigned either to category c_i or to its complement $\bar{c}_i$. An algorithm for binary classification can also be used for multi-label classification, because the problem of multi-label classification into k classes $C = \{c_1, ..., c_k\}$ can be transformed to the k independent problems of binary classification into two classes $\{c_i, \bar{c}_i\}$ for $i = 1, 2, \ldots, k$.

B. Training set, test set and cross-validation

The machine learning relies on the availability of an initial corpus $\Omega = \{d_1, d_2, \ldots, d_l\} \subseteq D$ of documents preclassified under $\{c_1, c_2, \ldots, c_k\}$. That means that values of the function $\widetilde{\Phi} : D \times C \to \{0, 1\}$ are known for every pair $(d_j, c_i) \in \Omega \times C$. A document d_j is a **positive example** of category c_i if $\widetilde{\Phi}(d_j, c_i) = 1$, and **negative example** of category c_i if $\widetilde{\Phi}(d_j, c_i) = 0$. Once the classifier Φ has been built it is desirable to evaluate its effectiveness. In this case, prior to the classifier construction, the initial corpus is split into two sets: a **training set** and a **test set**. The classifier Φ is built by observing the characteristics of training set documents. The purpose of the test set is to test the effectiveness of the built classifier. For each document d_j contained in the test set the classifier decision $\Phi(d_j, c_i)$ is compared with the expert decision $\widetilde{\Phi}(d_j, c_i)$. A measure of classification effectiveness is based on how often the $\Phi(d_j, c_i)$ value matches the $\widetilde{\Phi}(d_j, c_i)$ value for all documents in the test set. In evaluating the effectiveness of document classification by a certain classificator ***t*-fold cross-validation** [9] is a common aproach. The procedure of t-fold cross-validation assures statistical reliability of evaluating results. In this procedure t different classifiers $\Phi_1, \ldots, \Phi_t$ are built by partitioning the initial corpus Ω into t disjoint sets $\Omega_1, \ldots, \Omega_t$ and procedure of learning and testing is applied t times using Ω_i as a test set and $\Omega - \Omega_i$ as a training set.

The final effectiveness is obtained by individual computing the effectiveness of classifiers $\Phi_1, \ldots, \Phi_t$ and then averaging the individual results.

C. Measures of evaluation

We will mention only the most common measures that will be used for evaluation of classification for our experiment. Basic measures are **precision** and **recall**. Precision p is a proportion of documents predicted positive out of those that are actually positive. Recall r is defined as a proportion of positive documents out of those that are predicted positive. Usually there is trade off between precision and recall. That is why measures that take in account both of that two basic measures are good estimators of effectiveness. One of them is the F_β function [12] for $0 \le \beta \le +\infty$ defined as

$$F_\beta = \frac{(\beta^2 + 1) pr}{\beta^2 p + r} \quad . \qquad (1)$$

Here β may be seen as the relative degree of importance attributed to p and r. If $\beta = 0$ then F_β coincide with p, while for $\beta = +\infty$ F_β coincides with r. Usually, a value β=1 is used, which attributes equal importance to p and r.

D. Support vector machines

Support vector machine [4,6] is a classification algorithm that finds a hyperplane which separates positive and negative training examples with maximum possible margin. This means that the distance between the hyperplane and the corresponding closest positive and negative examples is maximized. The hyperplane is determinated by only a small set of training examples which is made up of closest positive and negative training examples. These examples are called the **support vectors**. A classifier of the form $f(x) = sign(w \cdot x + b)$ is learned, where w is the weight vector or normal vector to the hyperplane, b is the bias, and x is vector representation of the test document. If $f(x) = 1$, then document represented by vector x is predicted to be positive, while $f(x) = -1$ predicts that document is negative.

IV. EXPERIMENT

A. Data description

Our experiment has been conducted on the hierarchy of Web sites www.hr applied to the administrator until November, 2003. Hierarchy is represented by xml file provided by the search engine, and extended xml file created from original one by our software. For example, in the original xml file a site is represented in the following way:

<link id="**L6**" numVisits="**146**" dateAdded="**2003-aug-08**" rating="**1.80**">
Hrvatska - Informativni prirucnik
<desc>**Informativni prirucnik o Hrvatskoj na njemačkom jeziku.**</desc> .

Here we can see that in the original xml file, for each site, information are provided about its identity number, number of visits, data added, rating, referent link, name of the site and description of the site provided by the author. Such a representation is settled in the category which it belongs to. In the extended xml file only the description of the site is changed in a way that the text contained on the first level of links associated to the site is parsed. Our task is to compare classification effectiveness for these two different representations. Catalogue contains 602 categories and 12020 sites. We had categorized sites in the 14 topmost categories: About Croatia (AboutCro), Art & Culture (Arts), Business & Economy (Business), Computers & Networking (Comp), Education (Edu), Entertainment (Entert), Events (Events), News & Media (News), Organizations & Associations (Organiz), Law & Politics (Politics), Science & Research (Science), Society (Society), Sport & Recreation (Sports), and Tourism & Travelling (Tourism). The list of index terms for both representations is created by extracting all the terms present in the name of the site and its description, by discarding all the terms present in less then 4 sites and all the terms on the list of the stop words for Croatian language. By that procedure we got a list of 7239 index terms for original and of 18518 index terms for extended xml file. For the bag of words representation we have used TF-IDF weighting formula and we have normalized columns of the term-document matrix.

B. Results

For evaluation we have used 5-fold cross-validation procedure. The effectiveness of classification is evaluated by measures of precision, recall and F_1. The results are summarized in Table 1 and 2 and on Figures 1, 2 and 3. In the last row of Table 1 macroaverage of precision and recall is presented, while in the last row of Table 2 macroaverage of F_1 measure is presented. Macroaverage of the precision and recall is calculated simply as a mean values of precision and recall through the categories, while macroaverage of the F_1 measure is calculated by using formula (1) and macroaveraged values of precision and recall. We can see that overall results are the worst for representation by extended xml file. Macroaveraged precision for representation by extended xml file is by 3% lower then for representation by original xml file, while macroaveraged recall is by 10% lower for the extended representation. As a consequence of the lower results of macroaveraged precision and recall we have a 9% lower value of macroaveraged F_1 measure. Generally, results of precision are satisfactory for both representations, but the results of recall are not. Categories Events, News & Media , and Organizations & Associations have very low recall, but we did expect that this could happen, because sites that belong to that category may not contain key words for it. For example the site contained in category News & Media may not contain key words radio, magazine ... We did not expect low values of recall for categories Law & Politics and Science & Research. Recall for category Science & Research has increased after the extension of xml file. The reason for that may be that description of the sites provided by their owners doesn't contain key words after which category could be recognizable.

V. CONCLUSION

This research is just a beginning of a more comprehensive work on automatic categorisation of Croatian Web sites in hierarchical structure. We saw that extension of xml file did not result with more effective categorisation. Anyway, in the real situation when sites are crawled and not applied by the owners, description of the site is not provided by the owner. Our task is to find a model which will approximate manual indexing, in this particular case, as good as possible. Lots of improvements could be done. For example, in creating the list of indexing terms we did not apply stemming or lematization. This is transformation of the word on its root or basic form. Further, when dealing with text mining we are confronted with the problem of synonyms (more words with the same meaning) and polysemy (a word with more meanings). Bag of words is just a basic model and there is lots of improvements of that model that tend to overcome the problem of synonyms and polysemy. Application of some of those models is the subject of our further work.

Category	Original		Extended	
	Precision	**Recall**	**Precision**	**Recall**
AboutCro	90,26 ± 1,36	74,96 ± 1,15	90,15 ± 1,09	66,69 ± 1,58
Arts	78,64 ± 1,80	53,20 ± 2,02	79,77 ± 1,85	36,11 ± 1,45
Business	88,69 ± 0,60	88,09 ± 1,07	83,55 ± 0,54	87,93 ± 0,79
Comp	81,09 ± 0,96	59,41 ± 1,34	79,58 ± 0,36	41,72 ±3,71
Edu	85,79 ± 2,41	60,95 ± 5,74	84,47 ± 3,95	53,20 ± 7,31
Entert	79,10 ± 4,10	40,69 ± 3,77	81,02 ± 3,32	27,77 ± 2,88
Events	66,27 ± 8,94	19,70 ± 4,01	70,23 ± 18,33	12,88 ± 4,97
News	62,73 ± 8,38	27,43 ± 0,73	64,36 ± 1,07	15,86 ± 2,85
Organiz	62,54 ± 5,81	32,95 ± 1,77	57,45 ± 8,15	17,83 ± 3,73
Politics	73,48 ± 9,40	26,02 ± 3,99	60,35 ± 7,90	17,13 ± 6,08
Science	75,86 ± 4,09	31,42 ± 1,72	81,48 ± 5,21	30,08 ± 2,25
Society	77,86 ± 1,02	47,32 ± 2,80	75,35 ± 1,66	34,16 ± 1,82
Sports	86,30 ± 4,35	57,55 ± 2,96	83,98 ± 1,87	44,81 ± 2,23
Tourism	91,17 ± 1,21	77,76 ± 1,10	90,78 ± 1,34	69,15 ± 1,57
Macroaverage	78,56	49,82	75,60	39,67

Table 1. Precision and recall for 14 topmost categories of the original and extended representation of the hierarchy www.hr. In the last row macroaverage of these measures is presented.

Category	Original	Extended
AboutCro	81,89 ± 0,21	76,66 ± 1,26
Arts	63,46 ± 1,86	49,71 ± 1,68
Business	88,39 ± 0,32	85,68 ± 0,64
Comp	68,58 ± 1,22	54,67 ± 3,18
Edu	71,17 ± 4,26	65,12 ± 5,93
Entert	53,67 ± 3,58	41,27 ± 3,03
Events	29,96 ± 3,60	21,50 ±7,70
News	38,03 ± 0,74	25,37 ± 3,68
Organiz	43,03 ± 0,86	27,15 ± 5,03
Politics	38,35 ± 5,18	26,35 ± 8,02
Science	44,43 ± 2,25	43,85 ± 2,02
Society	58,81 ± 1,86	47,00 ± 2,04
Sports	69,03 ± 3,15	58,39 ± 1,41
Tourism	83,92 ± 0,42	78,50 ± 1,45
Macroaverage	60,97	52,03

Table 2. F_1 measure of topmost categories of the original and extended representation of the hierarchy www.hr. In the last row macroaverage of F_1 measure is presented.

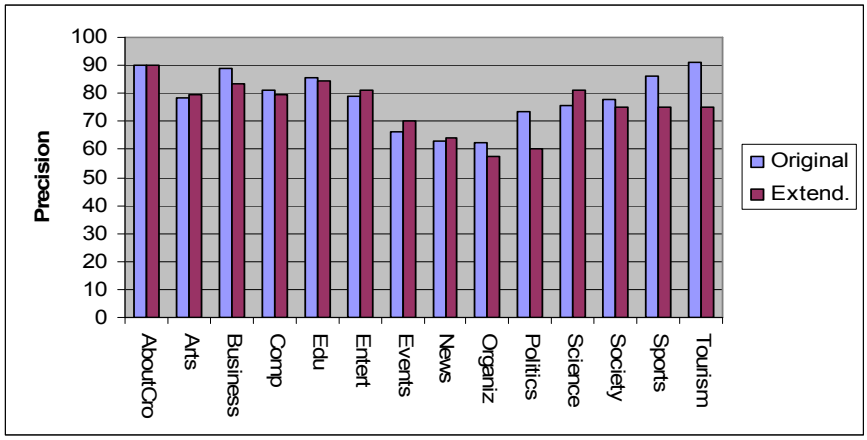

Figure 1. Precision for 14 topmost categories of the original and extended representation of the hierarchy www.hr.

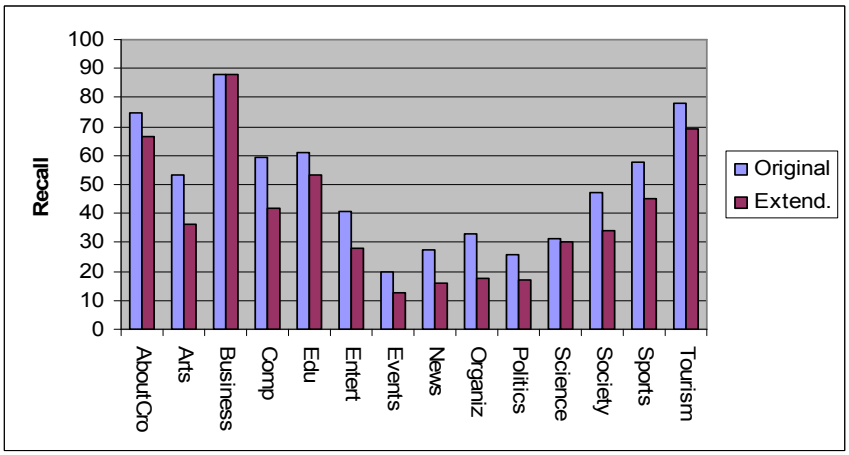

Figure 2. Recall for 14 topmost categories of the original and extended representation of the hierarchy www.hr.

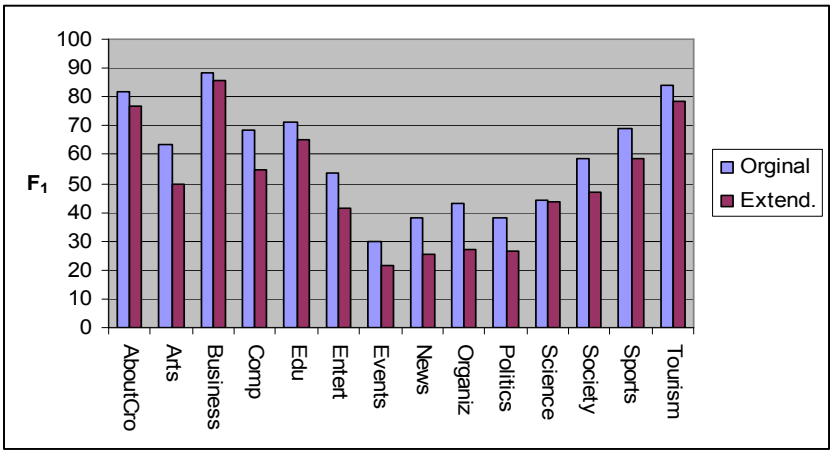

Figure 3. F_1 measure of topmost categories of the original and extended representation of the hierarchy www.hr.

REFERENCES

[1] G. Attardi, A. Gulli, F. Sebastiani. Automatic Web page categorization by link and context analysis, In *Proceedings of THAI-99, European Szmposium on Telematics, Hypermedia, and Artificial Intelligence,* Varese, IT, p. 105, 1999.

[2] R. Baeza-Yates, B.Ribeiro-Neto. *Modern Information Retrieval*, Addison-Wesley, ACM Press, New York, 1999.

[3] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, J. Kleinbrg. Automatic resource compilation by analiying hyperlink structure and associated text, *Computer Networks and ISDN Systems,* 30, p. 65, 1998.

[4] N. Cristianini, J. Shave-Taylor, *Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.

[5] N. Fuhr, N. Gövert, M. Lalmas, F. Sebastiani. Categorisation tool: Final prototype, Deliverable 4.3, Project LEA-8303 "EURUSEARCH", Commission of the European Communities, 1999.

[6] T. Joachims. Text categorization with support vector machines: Learning with many relevant features, In *Proceedings of the European Conference on Machine Learning,* Berlin, Springer, p. 137, 1998.

[7] D. Koller, M. Sahami. Hierarchically classifying documents using very few words. In *Proceedings of ICML-97, 14th International Conference on Machine Learning,* Nashville, US, p. 170, 1997.

[8] L.S. Larkey, W.B. Croft. Combining classifiers in text categorization, In *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval,* Zürich, CH, p. 289, 1996.

[9] T.M. Mitchell, *Machine Learning,* The McGraw-Hill Companies, Inc., New York, 1997.

[10] D. Mladenić, M. Grobelnik, Feature selection on hierarchy of web documents. *Decision Support Systems*, Vol. 35, No. 1,p. 45, 2003

[11] D. Mladenić. Turning YAHOO! Into an automatic Web page classifier. In *Proceedings of ECAI-98, 13th Ezropean Conference on Artificial Intelligence,* Bringhton, UK, p. 473, 1998.

[12] C. J. van Rijsbergen, *Information Retrieval,* 2nd edition, Butterworths, London, 1979.

[13] G.Salton, C.Buckley, Term-weighting approaches in automatic text retrieval, *Information Processing & Management,* Vol. 24, No. 5, p. 513, 1988.

[14] F.Sebastiani, Machine learning in automatated text categorization, *ACM Computing Surveys,* Vol. 34, No. 1, p. 1, 2002.