Using Salient Words to Perform Categorization of Web Sites*

Marek Trabalka and Mária Bieliková

Department of Computer Science and Engineering
Slovak University of Technology
Ilkovičova 3, 812 19 Bratislava, Slovakia
E-mail: trabalka@webinventia.sk, bielik@elf.stuba.sk
WWW: http://www.dcs.elf.stuba.sk/~bielik

Abstract. In this paper we focus on web sites categorization. We compare some quantitative characteristics of existing web directories, analyze the vocabulary used in descriptions of the web sites in Yahoo web directory and propose an approach to automatically categorize web sites. Our approach is based on the novel concept of salient words. Two realizations of the proposed concept are experimentally evaluated. The former uses words typical for just one category, while the latter uses words typical for several categories. Results show that there is a limitation of using single vocabulary based method to properly categorize highly heterogeneous spaces as the World Wide Web.

1 Introduction

Huge amount of web sites existing nowadays evolves a special type of web sites used for reference purpose – web directories. Web directories include links to other web sites together with a short description of their content. Web sites descriptions and corresponding links are stored in a hierarchy of categories. Hierarchies are usually defined by human maintainers. Usually only incremental additions are performed. Existing structure is rewritten very rarely.

Usefulness of web directories is similar to the Yellow pages. When a user is looking for an information or service, he simply browses through relevant categories in order to find matching web sites. However, manual creation and maintenance of the directory is quite expensive.

The aim of this paper is to present an approach to perform addition of new web sites into existing categorization automatically. We use results from the vocabulary analysis of established categorization hierarchy. The analysis is based on the novel concept of *salient words*, which is experimentally evaluated within a real collection of web sites.

There exist significant amount of work related to the categorization of documents. Many authors use for evaluation non-web texts like the Reuters corpus, medical OHSUMED collection or patent corpuses [11,5,9]. In fact, these collections are incomparable to a web site collection. Web site collections are extremely diverse in means of topic diversity, length of documents and variability of documents quality.

* This work was partially supported by Slovak Science Grant Agency, grant No. G1/7611/20.

Petr Sojka, Ivan Kopeček and Karel Pala (Eds.): TSD 2002, LNAI 2448, pp. 65–72, 2002.

In a past five years interest in web categorization of web documents rapidly grows. Most of existing approaches use existing web directories as a source of training and testing data [4]. Some authors just apply standard classification techniques to flattened categories [3]. Koller and Sahami [6] present an improvement of categorization speed and accuracy by utilizing hierarchical topic structure. They proposed small independent classifiers for every category instead of one large classifier for the whole topic set. Unfortunately, evaluations were done only on quite limited hierarchy of topics [2,6]. We performed broader analysis in order to find limitations of simple vocabulary analysis for detailed categorization.

The rest of the paper is organized as follows. In Section 2 we analyze structural characteristics of web directories. The analysis provides basis for proposed method of vocabulary analysis (Section 3). The concept of salient words is realized using words typical for one category and using words typical for several categories. In Section 4 we provide results of experimental evaluation. The paper concludes with summary and possible directions of research.

2 Structural Characteristics of Web Directories

At the present time there exist many web directories. Some of them are global; some of them are limited to some extent. There are various local web directories with respect to the country or language used. Also various thematic web directories exist that try to map more in-depth some particular field of interest.

Structural characteristics of local web directories are in most cases similar to global ones. Table 1 gives a comparison of two global directories and three local Slovak web directories.

Table 1. Comparison of web directories.

Site	Yahoo	DMOZ	Zoznam	Atlas	SZM
Language	English	English and others	Slovak	Slovak	Slovak
All categories	372 343	397 504	864	1 213	372
Top level categories	14	21	14	12	14
Second level categories	353	539	249	280	169
Third level categories	3 789	6 199	424	622	170
Depth of hierarchy	16	14	5	5	6
Average length of category title	14.05	12.03	16.01	15.43	12.84
Total number of sites	1 656 429	2 912 282	22 266	20 314	11 256
Average number under category	8.85	8.60	26.16	17.85	34.42
Average length of site title	22.37	23.20	18.71	16.77	21.26
Average length of site description	67.55	96.21	72.42	111.22	69.43

We use Yahoo and DMOZ global web directories. *Yahoo* (`www.yahoo.com`) is the best-known commercial web directory existing since 1995. *DMOZ – Open Directory Project* (`www.dmoz.com`) is a non-commercial web directory updated by volunteers. *Zoznam* (`www.zoznam.sk`), *Atlas* (`www.atlas.sk`) and *Superzoznam* (`www.szm.sk`) are the three most popular Slovak web directories.

Web directories are generally quite similar each to other. They have many categories in common and also their look and feel is the same. The main difference between local and global web directories is in the number of covered web sites that affects also size of the hierarchy of categories. We explored also other web directories and found out that they share almost the same characteristics. The number of top-level categories is usually between 10 and 16; typical number of subcategories is between 2 and 30. Lengths of titles are also very similar in average. The only difference is sometimes in the length of site descriptions where some directories limit the maximum length.

3 Analysis of Vocabulary and Categorization

Existing web directories are the great source of information for training categorization. Most of them are manually checked and therefore their quality is high. Furthermore, they contain large amount of information that could be used to acquire explicit knowledge about the categories and also about the whole domain.

There is a strong correspondence between a category and the vocabulary used in web sites assigned to the category. We consider the following text categorization assumption: *it is possible to correctly assign a web site into the category only by means of its textual information.* In a real life this assumption is not always the truth, indeed. There exist web sites containing most of their content in images or other non-textual kind of presentation that prohibits categorization by analyzing only the text. Analysis of images is beyond the scope of our research, we assume that such information can be converted to the text.

The web directory covers internal information – stored directly in the web directory (URLs, site descriptions and title) and external information – the web sites themselves referred by URLs. We use only the internal information to build representative texts. Of course when categorizing a new web site into the hierarchy we have to deal with its content as the only available information. It is obvious that using also external data, i.e. the content itself, provides more valuable data. On the other hand, such approach would require more computing resources.

Text categorization assumption implies the possibility to create a classifier able to correctly classify web sites by examining their textual contents. It is necessary to have a model of every category to compare the web sites with. There were proposed various models in information retrieval community to deal with a document clustering that could be applied in our case as well (for review see [9]). Commonly used is the Vector Space Model proposed by Salton in SMART project [8]. In this model a feature vector represents every document, query or group of documents. Usually, features are words or stems, and their values in the vector correspond with the number of occurrences in the object. Similarity of objects is computed by cosine of angle between these two vectors:

$$r = \frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{i=1}^{n} q_i^2 \sum_{i=1}^{n} p_i^2}} = \frac{Q, D}{||Q||\,||D||} = \cos\theta$$

This method has several advantages, including easy implementation. Its main disadvantage is high computational cost due to high dimensionality of vectors. When words or their stems are used as features the vector could have dimensionality of tens or even hundreds of thousands that significantly slows down the comparison process.

Many approaches to improve this method focus primarily on dimensionality reduction of the feature vector [6]. Dimensionality reduction can be achieved by selection of the most useful words. (e.g., the words able to distinguish between categories). Figure 1 depicts the difference between common word *'and'* and a category specific word *'newspaper'*. The figure displays how differ relative occurrences of these two words in documents within top-level categories. General terms have similar relative occurrences in all categories while category specific words are often used in one or few categories and in others are quite rare.

Our approach is to explicitly find the words significant for a category distinction within neighboring categories. Such words are identified for every category and its respective direct subcategories because a word able to distinguish between subcategories of one category may have similar occurrences between subcategories of another category. We call such words *salient words*.

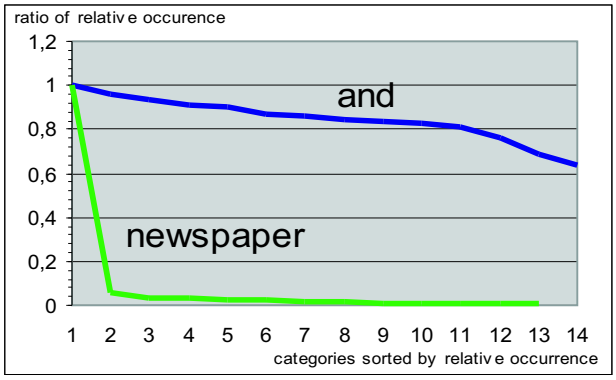

Fig. 1. Occurrence of common word and category specific word.

3.1 Categorization Using Words Typical for One Category

When human roughly analyses topic of a text he often relies on salient words – terms typical for the topic. Brief view on the article vocabulary without in-depth semantic analysis is often sufficient for human to distinguish the topics.

Similarly to human approach, we suggest a concept of salient words to be applied in automatic classification of web documents. Main idea is to use for categorization only the words typical for a particular category. This leads to a significant reduction of computation costs during the categorization. Together with hierarchical analysis it makes the processing fast and efficient.

The method of identifying salient words consists of the following steps performed for every category (including root category):

1. Collect words used in the category with a number of occurrences above defined threshold (to refuse rare words).
2. Perform steps 2a – 2c for every collected word:

(a) Compute the relative occurrence of the word w for every subcategory c, i.e. the number of word's occurrences divided by occurrences of all words within the subcategory.
(b) Compute the sum of the relative occurrences within direct subcategories $sum_{c,w}$.
(c) Find the maximum of relative occurrences within direct subcategories $\max_{c,w}$.
(d) If $\max_{c,w} \geq sum_{c,w} \times salient\_const$, mark the word as a salient for the subcategory with the highest $max_{c,w}$. Remember the strength of this word for a category distinction as $strength_{c,w} = \max_{c,w} / sum_{c,w}$. The $salient\_const$ constant is a tunable parameter usually between 0.5 and 1. The higher the constant is the less number of salient words we acquire.

Text categorization is performed by traversing through the hierarchy looking for the best matching category. The process starts by computing the relative occurrence for all words in given text. Next, for the root category the following steps are performed:

1. Find all direct subcategories of the category; terminate if there are no subcategories.
2. Compute similarity between every subcategory and a given text as
   $similarity_c = \sum_{i=salient\_word} strength_{c,j} \times relocc_i$
   where $relocc_i$ is the relative occurrence of the word i within given text.
3. Find the maximum of similarities $\max_{sim}$. If $\max_{sim} < similarity_{bias}$, terminate. If it is above the bias, append the subcategory with the $\max_{sim}$ in the resultant stack and perform recursively these steps for the subcategory.

3.2 Categorization Using Words Typical for Several Categories

Actually, there exist many significant words that are not typical just for one category but for two or more categories. If the presence of a word could eliminate at least few categories we call such word *separable*. Separable words include also salient words.

We collect separable words for every category. Separable words are related to a parent category. They are used to distinguish between direct subcategories of the category. The feature vectors for sibling categories and the list of words used in the feature vector are kept.

The method of identification separable words consists of the following steps performed on every category (including root category):

1. Collect words used in the category with the number of occurrences above defined threshold.
2. For every collected word compute its relative occurrences for every subcategory. If at least for one category value exceeds $separable_{bias}$, insert the word into the list of separable words for the category. For every subcategory insert the number of occurrences of this word into the feature vector of the subcategory.

For a given text of a web site the category tree is traversed and at each step the closest category feature vector is selected. The process starts by computing the number of occurence for all words in the given text. Next, for the root category the following steps are performed:

1. Find all direct subcategories of the category; terminate if there are no subcategories.

2. Prepare text feature vector as a list of occurrences of separable words for the examined category.
3. For every subcategory compute similarity between the subcategory feature vector and the text feature vector as

$$similarity_c = \frac{\sum_{i=1}^{n} cat_i \times doc_i}{\sqrt{\sum_{i=1}^{n} cat_i^2 \sum_{i=1}^{n} doc_i^2}}$$

   where cat_i, resp. doc_i is the number of occurrences of i-th separable word in the processed category, resp. the text of the document.
4. Find maximum of similarities $\max_{sim}$. If $\max_{sim} < similarity_{bias}$, terminate. If $\max_{sim}$ is above the bias, append the subcategory with $\max_{sim}$ in the resultant stack and perform recursively these steps for the subcategory.

4 Experimental Evaluation

We implemented proposed methods of web sites categorization and made several improvements and optimizations. Firstly, to improve both, speed and recall we employ stemming of words using the Porter's suffix removal algorithm [2].

To reduce number of different words to be analyzed we use approximately 500 stop words that were removed from all processed texts. We also removed stems with a rare occurrence in the whole web directory (bellow the threshold of 10 occurrences in our experiments). This decreased the number of stems from 299 470 to 29 201. We used the vocabulary of titles and short descriptions of web sites to acquire significant words. Analyzed Yahoo web directory contained almost 400 000 categories. Most of them contained only a few sites and therefore did not provide enough text for training. For the evaluation we selected only categories that contained at least 1 000 sites (including their subcategories). For acquired 978 categories we built the vocabulary from descriptions of sites registered within these categories and their subcategories.

We randomly selected web sites registered within the Yahoo and downloaded their contents up to 100 kB. Many researchers analyze only web site's first page directly referred by registered URL [5,3] or snippets returned by the search engine [2]. We decided to download larger portion of the web site in order to analyze whether increased amount of data will improve quality of the analysis. We experimented with two sets of the web sites. Smaller set A contains 369 web sites with more than 100 kB of text per site while the larger set B contains 1277 web sites with at least 10 kB of text per site.

We used the set A to analyze the impact of the size of analyzed portion of the web site on quality of results. We compared the results based on a starting page, first 1 kB of text, first 10 kB of text and first 100 kB of text. Table 2 shows the results of analysis using words typical for one category. We obtained the best results for 10 kB portion of a web site. It also proves our hypothesis that using only the first page for categorization is not always sufficient.

Then we used 10 kB parts of more than 1 000 web sites to analyze overall quality of proposed approach and dependence of estimation with respect to the appropriate category. Table 3 shows that there are significant differences between the categories. In-depth analysis of the most erroneous categories shows that many invalid top-level assignments were to Business & Economy and Computers & Internet categories.

Table 2. Analysis of web pages with different size.

Correctly estimated levels of categories	First page	1 kB	10 kB	100 kB
0 levels	55 %	46 %	42 %	43 %
1 level	11 %	8 %	8 %	9 %
2 levels	16 %	17 %	19 %	18 %
3 levels	9 %	11 %	10 %	13 %
4 levels	6 %	11 %	13 %	10 %
5 levels	1 %	3 %	2 %	2 %
6 levels	0 %	1 %	2 %	1 %

Table 3. Categorization of web sites according different categories.

Correctly estimated levels of categories	Overall	Arts	Regional	Business	Computers	Entertainment
0 levels	44 %	68 %	47 %	45 %	26 %	19 %
1 level	10 %	3 %	2 %	14 %	31 %	11 %
2 levels	18 %	4 %	15 %	22 %	19 %	22 %
3 levels	10 %	9 %	16 %	5 %	20 %	20 %
4 levels	9 %	5 %	7 %	5 %	1 %	18 %

5 Conclusion

In this paper we described two methods for categorization of web sites based on analysis of salient words. We use short descriptions of web sites in a web directory to select words useful to distinguish categories. Categorization process uses category tree to limit the number of necessary comparisons and speed up the processing. We evaluate success of categorization by comparing estimated and actual categories of the web site within web directory. We also show how size of downloaded portion of a web site affects the result of categorization and present the difference in success within different top-level categories.

In the further research we would like to compare results of our methods when trained on full texts of web sites rather than their short descriptions in the web directory. We also plan to extend the amount of evaluated web sites in order to gain more precise results.

References

1. Brin, S., Page, L.:The Anatomy of a Large-Scale Hypertextual Web Search Engine (1998).
2. Dumais, S., Chen, H.: Hierarchical Classification of Web Content. In: Proc. of 23rd Int. ACM Conf. on Research and Development in Information Retrieval (SIGIR), Athens, Greece (2000) 256–263.
3. Mase, H.: Experiments on Automatic Web Page Categorization for IR system. Technical report, Standford University (1998).
4. Mladenic, D.: Turning Yahoo into an Automatic Web-Page Classifier. In: Proceedings of ECAI - European Conference on Artificial Intelligence (1998).
5. Karypis, G., Han, E.: Fast Supervised Dimensionality Reduction Algorithm with Applications to Document Categorization & Retrieval (2000).

6. Koller, D. and Sahami, M., Hierarchically classifying documents using very few words, in International Conference on Machine Learning (ICML) (1997) 170–178.
7. Porter, M. F.: An Algorithm for Suffix Stripping. Program, 14 (3) (1980) 130–137.
8. Salton, G.: A New Comparison Between Conventional Indexing (MEDLARS) and Automatic Text Processing (SMART). In: Journal of the American Society for Information Science 23 (2) (1972) 75–84.
9. Trabalka, M.: Document Retrieval. A Written Part of Ph.D. Examination. Slovak University of Technology (2001).
10. Wang, K., Zhou, S., He, Y.: Hierarchical Classification of Real Life Documents. First SIAM International Conference on Data Mining (2001).
11. Yang, Y., Pedersen, J. O.: A Comparative Study on Feature Selection in Text Categorization. In: Proc. of 14th Int. Conf. on Machine Learning (1997).