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
trabalka@webinventia.sk, bielik@elf.stuba.sk

Abstract. In this paper we focus on categorization task for web sites. We compare some quantitative characteristics of existing web directories, analyze vocabulary used in descriptions of 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. Experimental evaluation compares two realizations of proposed concept. The former uses words typical for just category, while the latter uses words typical for one or few categories. Results show that there is a limitation of using single vocabulary based method to properly categorize such heterogeneous space as is the World Wide Web.

1 Introduction

Huge amount of web sites existing nowadays evolves a special type of web sites used for reference purposes. These are often called 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 fixed and defined by humans maintainers. Usually only incremental additions are performed. Existing structure is rewritten very rarely.

Usefulness of these 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 some matching web sites. Unfortunately, creation and maintenance of such directories is quite expensive because it is usually done manually by humans.

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 vocabulary analysis of established categorization hierarchy. The analysis is based on the novel concept of salient words, which is experimentally evaluated within a real web sites collection.

There exist significant amount of work related to categorization of documents. Many authors evaluate proposed approach on non-web texts like 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 differences, lengths of documents and variability of quality of documents.

In a past five years interest of web categorization rapidly grows. Most of existing approaches to categorize web documents use existing web directories as a source of training and testing data [4]. Some approaches just apply standard classification techniques to flattened categories [3]. Koller and Sahami [6] present an improvement of categorization speed and accuracy when utilizing hierarchical topic structure. They proposed small independent classifiers for every category instead of one large classifier for the whole topic set. Unfortunately, evaluations of this proposal 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 statistical and structural characteristics of web directories. This analysis provides basis for proposed method for analysis of vocabulary (Section 3). Concept of salient words is realized using words typical for one category (Subsection 3.1) and using words typical for more categories (Subsection 3.2). In Section 4 we provide results of experimental evaluation of the proposed approach. Paper concludes with summary and possible directions in this research.

2 Statistical and 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.

Local web directories are usually very similar to global ones, except they usually collect web sites written in a local language and cover only pages related to the same country or language. 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* (http://www.yahoo.com) is the best-known commercial web directory existing since 1995. *DMOZ* – Open Directory Project (http://www.dmoz.com) is a non-commercial web directory filled by volunteers. *Zoznam* (http://www.zoznam.sk), *Atlas* (http://www.atlas.sk) and *Superzoznam* (http://www.szm.sk) are the three most popular Slovak web directories.

Web directories are generally quite similar each to other. They have usually 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. 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 great source of information for computer learning process. Most of them are manually checked and therefore their quality is quite high. Furthermore, they contain large amount of information that could be used to acquire explicit knowledge about the categories and also about the whole covered domain. The web directory contains an internal information – stored directly in a web directory (URLs, site descriptions and title) and external information – web sites themselves referred by URLs.

We believe there is a strong correspondence between category and vocabulary used in websites assigned to particular 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 identity in images or other non-textual kind of presentation that prohibits categorization by analyzing only their text. Because analysis of images is beyond the scope of our research, we decided to ignore these web sites (or assume that such information can be converted to the text).

In order to deal with a vocabulary related to particular category we have primarily two options. We can use only the internal information to build representative texts, i.e., short titles and descriptions of web sites. Much more resource consuming option is to acquire also external data, i.e., download referenced web sites or at least their parts. It is obvious that the second approach provides more valuable data. On the other hand, it would be much more complex to realize and would require more computing resources. We proposed an approach, which uses for vocabulary analysis only internal web site descriptions. Of course, when we categorize new web site into the hierarchy, we have to deal with its real contents actually.

Text categorization assumption implies the possibility to create a classifier able to correctly classify web sites by examining their textual contents. To do this, it is

necessary to have a model of every category to compare 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]). Probably the most used is the Vector Space Model (VSM) proposed by Salton in SMART project [6]. In this model a feature vector represents every document, query or group. Usually, features are words or stems, and their values in vectors correspond with number of occurrences in the object.

Similarity of objects is computed by cosine of angle between those two vectors:

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

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

Therefore, 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. We are looking for 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.

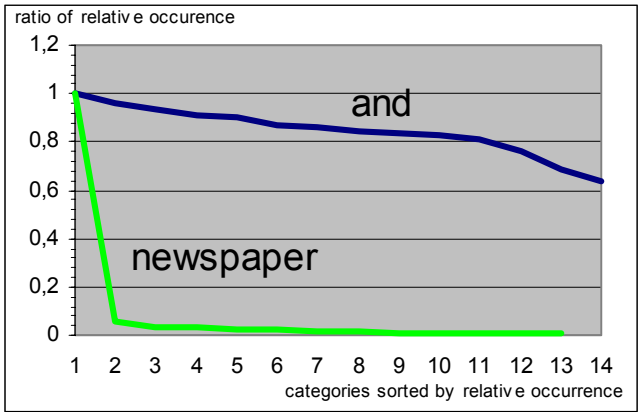

Fig. 1. Occurrence of common word and category specific word.

Our approach is to explicitly find the words significant for a category distinction within neighboring categories. Such words should be identified for every category and its respective direct subcategories, because word able to distinguish between

subcategories of one category may have similar occurrences between subcategories of another category. We call such words able to distinguish categories *salient words*.

3.1 Categorization of Web Sites Using Words Typical for One Category

When human roughly analyses topic of some text, he often relies on salient words – special terms typical for some topics. Brief view on the vocabulary in an article without in-depth analysis of sentences is often sufficient to distinguish its themes.

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

Process of identifying salient words consists of performing the following steps for every category including root category:

1. Find all direct subcategories of the category. End if there are no subcategories.
2. Collect words used in category with number of occurrences above defined threshold (i.e., rare words are not considered).
3. Perform steps 3.1 – 3.3 for every collected word
   3.1. Compute *relative occurrence* of word for every subcategory. Relative occurrence is the number of word's occurrences divided by occurrences of all words within the category.
   3.2. Compute sum of computed relative occurrences within direct subcategories. Let it be $sum_{cat,word}$.
   3.3. Find maximum of relative occurrences within direct subcategories. Let it be $max_{cat,word}$.
   3.4. If $max_{cat,sword} \geq sum_{cat,word} * salient\_constant$, mark the word as salient for the subcategory with this highest $max_{cat,word}$. Remember strength of this word for a category distinction as $strength_{cat,word} = max_{cat,word} / sum_{cat,word}$. Constant *salient_constant* is a tunable parameter usually between 0.5 and 1. The higher the parameter is, the less number of salient words we get.

After collecting salient words, process of text categorization is quite straightforward. It traverses through the hierarchy looking for the best matching category:

1. Compute relative occurrences for all words in a given text.
2. For the root category perform steps 3, 4 and 5.
3. Find all direct subcategories of the category. End if there are no subcategories.
4. Compute similarity between every subcategory and a given text as

$$similarity_{cat} = \sum_{i=salientword} strength_{cat,i} * relocc_i$$

   where $relocc_i$ is the relative occurrence of word *i* within given text.
5. Find maximum of similarities. If it is less then *similarity_bias*, return. If it is above bias, append the subcategory with this maximal similarity in the result stack and perform recursively steps 3, 4 and 5 with this category.

3.2 Categorization of Web Sites Using Words Typical for More Categories

Actually, there exist many significant words that are not typical just for one category, but for two or more categories. If a 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. These words are used to distinguish between some of its direct subcategories. Then we hierarchically use the vector space model to find the best matching category for given text.

Unlike salient words for a category, separable words are related to a parent category. We store feature vectors for sibling categories and list of words used in this vector. Then we traverse the category tree looking for the category vector closest to document vector.

Process of identification of separable words consists of performing the following steps on every category including root category:

1. Find all direct subcategories of the category. End if there are no subcategories.
2. Collect words used in category with number of occurrences above defined threshold.
3. 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 list of separable words for the category. For every subcategory insert number of occurrences of this word into feature vector of subcategory.

For given text of web site we traverse category tree and choose at each step the closest category feature vector:

1. Acquire occurrences for all words in the given text.
2. For root category perform steps 3, 4 and 5.
3. Find all direct subcategories of the category. End if there are no subcategories.
4. Prepare text feature vector as a list of occurrences for separable words for examined category.
5. For every subcategory compute similarity between subcategory feature vector and text feature vector as

$$similarity_{cat} = \frac{\sum_{i=1}^{n} cat_{cat,i} doc_i}{\sqrt{\sum_{i=1}^{n} cat_{cat,ii}^2 \sum_{i=1}^{n} doc_i^2}}$$

   where $cat_{cat,i}$ resp. doc_i is the number of occurrences of i-th separable word in a category *cat* resp. text of the document.
6. Find maximum of similarities. If it is less then *similarity_bias*, end. If it is above bias, append the subcategory with this maximal similarity in result stack and perform recursively steps 3 to 6 with this category.

4 Experimental Evaluation

We realized proposed approach and made several improvements and optimizations to proposed methods. Firstly, to improve both, speed and recall we employ stemming of words. We use simple Porter's suffix removal algorithm [2]. We have tested Lovin's algorithm as well, but its results were worse, therefore we chose Porter's.

In order to deal with high number of different words to be analyzed, we removed two groups of words. Firstly, we use list of approximately 500 stop words that were removed from all processed texts. Then we also removed rare stems found less than threshold (10 in our experiments) in the whole web directory. This decreased the number of stems from 299 470 to just 29 201.

We used vocabulary of titles and short descriptions of web sites to acquire significant words. Analyzed Yahoo web directory contains almost 400 000 categories. Most of them have only a few sites and therefore not enough text for training. For testing purposes we chose only those categories containing at least 1 000 sites (including those within their subcategories). For acquired 978 categories we built vocabulary from descriptions of all sites registered within a category and all of its subcategories.

We chose random web sites registered within 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 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 analysis. For our analysis we use two sets of 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 more than 10 kB of text per site.

We use set *A* to analyze how much size of analyzed portion of the website influences quality of the results. We compared results given by the analysis of 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. The results show the best values for 10 kB portion of a web site. It proves our hypothesis that using only a first page for categorization is not sufficient.

Table 2. Analysis of web pages with different size.

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

As the set *A* shows 10 kB portion of text optimal for categorization, we use 10 kB portions of more than 1 000 web sites to analyze overall quality of suggested approach and dependence of estimation with respect to the correct category. As Table 3 shows, there are significant differences between categories. More in-depth analysis of the most erroneous categories shows that many invalid top-level assignments lead into Business & Economy and Computers & Internet categories.

Table 3. Categorization of web sites according different categories.

Correctly estimated	Overall	Arts	Regional	Business	Computers	Entertainment
0 category	44%	68%	47%	45%	26%	19%
1 category	10%	3%	2%	14%	31%	11%
2 categories	18%	4%	15%	22%	19%	22%
3 categories	10%	9%	16%	5%	20%	20%
4 categories	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 web site affects the result of categorization and present the difference in success within different top-level categories.

In 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 web directory. We also plan to extend 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 PhD 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).