Classification of e-commerce websites by product categories

George Moiseev

Higher School of Economics, Moscow, Russia
gvmoiseev@edu.hse.ru

Abstract. Nowadays, the number of e-commerce websites steadily grows. Therefore, it is hard to collect and analyze such websites manually. Meanwhile, there are many market researchers and aggregation services that need to collect e-commerce websites for some reasons, for instance find them in a predefined domain zone or find only sites that belong to a certain product category. This paper proposes several methods for improving the preprocessing and the feature extraction stages of the web sites classification process. They are applied to the task of e-commerce websites automatic classification based on the sold product type. Experimental results show that proposed methods improve the classification accuracy.

Keywords: e-commerce website classification, product classification, web-mining, web page classification

1 Introduction and Related Works

One of the most common problems nowadays is a high amount of mostly unorganized information on the web. With the exponential growth of data around the web the arrangement of the information becomes an important task for assisting users and companies in storing and retrieving the information. One example of such task is an automated e-commerce websites categorization problem which also includes the issue of retrieving such sites from the web. This problem comes from a high need of clustered or categorized e-commerce websites for market researchers who need to take into account different types of statistics in the e-commerce sphere, such as [1], comparison of shopping engines like Google Shopping or Yandex Market [2], information retrieving systems [3] followed by other types of services. Normally, categorization by sold product type attracts the most interest.

To complete the description of the raised problem we have to specify that by *«e-commerce website»* we mean only business-to-customer (selling consumer goods and/or services to customers to earn a profit) or business-to-business (one business makes a commercial transaction with another) online shopping and we do not include customer-to-customer type of shopping when customers interact with each other while business only facilitates an environment. This makes the e-shops retrieving process more complex.

At first sight, the raised problem is a part of a wider text categorization issue or a web page classification task and it can be solved by direct borrowing of existing algorithms from machine learning literature dedicated to these issues [4-7]. Nevertheless, the solution is far from being so straightforward. Web pages are highly structured and filled with noisy content such as javascript code, advertisements and copyrights. Without taking these factors into account, they would have negative impact on performance of pure text classification algorithm. It has been proved that exploitation of the structure of a web page (HTML tags, hyperlinks) enhances the quality of classification [12].

Although most web page classification [6-8] algorithms apply noise reducing techniques and use structure of web pages to improve classification, there are still issues to discuss and ways to improve.

Firstly, classifying a website raises some ambiguous questions: which webpages of the website should be downloaded and processed, should hyperlinks from the main page be used and how, is the content from the main page more valuable than content from other pages, should one consider content from all pages of the website in feature selection process.

The second non trivial question is related to the use of a web page structure in classification process. Does one need to take into account words location on the web page and how to perform that? Should display properties of the content be considered in classification process?

Another significant point to consider is a language-based approach. Because of particularly high interest in research on the Russian online market we mostly focus on classifying Russian websites. Most researches on this topic study only web pages in English or try to develop language-independent method [7]. Concentrating on a single language area of the web allows us to use some language specific features that enhance the quality of classification. An interesting example of such features is using transliterated words which are frequently occurred in markup tags or hyperlinks on typical e-shop webpage.

Similarly, narrowing the sample of categorized websites to online stores gives us the opportunity to exploit the domain knowledge: predefine some handcrafted features, use typical e-shop webpage structure, check the existence of special HTML tags and look for some specific words in hyperlinks. Exploiting most of these features is impossible without restricting the category of websites.

In this paper, we propose an approach to download and preprocess a website and feature engineering techniques concerned with the mentioned issues.

The rest of paper is organized as follows: downloading and preprocessing stage is described in Section 2, feature engineering methods are discussed in Section 3 and Section 4 contains experiment environment, description of the dataset and experiments results. Finally, we conclude our work and point out the related future work.

2 Downloading and preprocessing

Most e-commerce websites consist not only of a home (main) page, but also of several additional pages such as «shopping cart» page, «catalogue» page or «help» page. However, most website catalogues and database don't store links to all pages of websites. Usually, only the home page link of a website is saved.

Therefore, the expected input of the system is suggested to be a list of home page links. Certainly, a classification process may be based only on features extracted from the main page. But there are cases when the main page may contain many images or Flash object and less textual content. Some researches on this topic show that exploiting information from hyperlinks improves the accuracy of classification [10-11]. Furthermore, it is a common practice for e-commerce websites to place information on sold products on individual pages or on a page with catalogue. Since this information is principal for our categorization we need find a way to gather and exploit the data from some pages of a given website.

The primary way to obtain other webpages when one has only the home page is to use hyperlinks located on the home page. The issue in this case is that we have to retrieve only useful hyperlinks because most hyperlinks contain useless or even deteriorative data for classification. In the course of the study we have derived an empirical rule which consists of ignoring links that satisfy one or more conditions listed below:

1. the link refers to a different website;
2. the hyperlink anchor text contains terms frequently used in anchor texts of web pages from different categories. The list of these terms is a union of sets of the most frequent anchor texts from each category. Some examples of such terms are: «доставка» (delivery), «контакты» (contacts), «помощь» (help) and «корзина» (shopping cart).

Another issue here is the exploitation of the content from selected links. Several researches on this topic show that this should be done very carefully. The obvious method here is to concatenate text from the main page and texts from other pages, and afterwards use concatenated text as input for classifier. But Chukrabarti in his research shows that this method performs dismally, and classification without concatenating is more efficient [9]. In the course of the study we carried out a similar experiment and observed same results in binary classification task. Our idea is to extract only meta tags and title (which usually gives a good summary of the page) from other pages of the website and combine them with information from the main page. This approach eliminates the most possible noise from other pages by extracting only summarized information from meta tags and title. But in cases when the possibility of topic drift and noisy information is quite low using entire pages may have its advantage because it allows considering more information about the web site. Thus both variants are tested and compared with classifying only by main page. Experimental results can be found in Section 4.

The next preprocessing steps are quite obvious: removing noisy content such as copyright, advertisements or javascript code, removing stop words, extracting pure text from the page, tokenization, lowercase conversion and stemming. Stemming process uses Snowball algorithm. To detect advertisements we use separate classifier with the set of features including occurrences of links or words referring to widely used online advertisements systems («ad.yandex.ru», «adsense», «adserver», «adsystem», «adsale», «openx»), number of links in the tag, number of words, the proportion of capital letters, proportion of full stops, links and tag density, font size, margins. To remove copyright we check the occurrences of word «copyright» or «©» symbol and date. However, the initial web page with markup tags is saved for the feature extraction process.

3 Feature Engineering

As it was mentioned before, HTML tags provide significant information about the content of a web page. For instance, words nested in <title> tag or <description> meta tag are usually more important for classification than words from <body> as they should give a summary of the page. Also, authors of a web page use header tags, color or font to emphasize some information. There are special tags for important text such as <strong> and <em>. But the style of web pages can be very different: some authors use special tags to emphasize several important words on a whole page while others may mark every second sentence.

Our idea is to weight terms against the nearest tag they are nested in and to calculate the weight of tags inversely proportional to their frequency (i.e. the more frequent the tag is, the less valuable enclosed terms are).

The term weighting formula for the *i*th term in the *k*th web site is derived from TF-IDF [15] as follows:

$$W_{ik} = \frac{tf_{ik}\log\frac{N}{n_i}}{\sqrt{\sum_{j=1}^{N}(tf_{ij}\log\frac{N}{n_j})^2}} \tag{1}$$

where n_i is the number of websites where the *i*th term appears, N – total number of web sites in the sample and tf_{ik} is computed as:

$$tf_{ik} = \sum_t^T w(t)\mathrm{f}(i,k,t) \tag{2}$$

where T is the set of all tags of *k*th web site, f(*i*, *k*, *t*) is the frequency of the *i*th term in tag *t* from web site *k* and $w(t)$ is calculated as follows:

$$w(t) = \frac{1}{\sum_x^T[x=t]} \tag{3}$$

Besides these features binary *«e-commerce or not»* classification process considers some handcrafted empirical binary features which were discovered through careful analysis of several tens of typical e-commerce web sites. These binary features are checked before the stemming stage because in some of them the form of a word is important. They are presented in the table below in common regular expression notation:

Table 1. Handcrafted features for e-commerce web sites classification

Feature	Remarks
корзин[а-я]	shopping cart
[a-z]*cart	often occurs in shopping cart hyperlink anchor text
[a-z]*basket	often occurs in shopping cart hyperlink anchor text
достав[илк][а-я]*	delivery
самовывоз[а-я]*	pickup
ассортимент[а-я]	variety
([0-9]*|)руб	indicates price
сумм[ауые]	total cost
товар[а-я]*	good
оплат[а-я]*	payment
заказ[а-я]*	order
купить	to bye
покуп[а-я]*	purchase
pay(ment|)	often occurs in payment hyperlink anchor text
pric(i|e)[a-z]*	often occurs in price list hyperlink anchor text
наличи[а-я]	presence
(розниц[а-я]|розничн[а-я]*)	retail
скидк[а-я]	discount
цен([аеоуы].{0,2}|ник)	indicates price
аксессуар[а-я]*	
(рас|)продаж[а-я]{0,2}	sale
products?	often occurs in catalogue hyperlink anchor text
интернет.{0,5}магазин.{0,5}	online shop
delivery[a-z]*	often occurs in delivery hyperlink anchor text
sales?	often occurs in catalogue hyperlink anchor text
оптом	wholesale
oplat(a|y)	often occurs in payment hyperlink anchor text
bitrix	often occurs at web sites about creating e-shop sites.

buy	often occurs in purchasing hyperlink anchor text
compare	
dostavka	often occurs in delivery hyperlink anchor text
quantity	often occurs in tags about some goods and their available quantity
каталог	catalogue
подар[а-я]*	gift
гаранти.	guarantee

4 Experiments

In order to test the effectiveness of using tags in feature engineering and considering meta tags and title from pages obtained via hyperlinks, several experiments are conducted. Since the focus of this paper is on the preprocessing and feature engineering stages, we choose one of the most popular classifiers – Support Vector Machine (SVM). This powerful learning algorithm was proposed by V. Vapnik [13] and has been proved as one of the most powerful algorithms for text categorization [14].

4.1 Dataset

The dataset was received from datainsight.ru and completely consists of websites in Russian. General dataset consists of two subsets: one for binary *«e-commerce or not»* classification task and the second one for categorization. Both subsets were gathered separately. Thus not all web sites from second subset are presented in the first subset and vice versa.

The dataset for binary classification contains 1312 e-commerce and 1077 non e-commerce web sites. Some of non e-commerce web sites are specially chosen C2C sites while others are chosen randomly from .ru and .рф domains. The dataset for categorization contains 1448 web sites in total. The list of available categories and numbers of web sites in each of them are listed in the Table 2. General department stores like amazon.com or aliexpress.com belong to the «General stores» category.

Table 2. E-commerce product categories dataset

category id	category name	number of web sites
0	Auto products	138
1	Medical goods	289
2	Health and beauty products	114
3	Appliances and electronics	168
4	Household goods	171
5	Furniture	79
6	Souvenirs, presents	36

7	Media (books, disks and concert tickets)	69
8	Jewellery and clocks	43
9	Technical and industrial equipment	65
10	Food and kindred products	72
11	Pet supplies	44
12	Sport equipment and hobbies	51
13	Clothing and footwear	65
14	General stores	44

All data is presented in the following format (with examples from the second subset):

Table 3. Example of how the data set is stored

domain name	category id
seving.ru	9
evalar.ru	1
mojon.ru	13
hunt.ru	12
…	…

4.2 Evaluation

We use common measures to evaluate the performance of the classifier: precision, recall and F-score[16]. To evaluate the average performance between multiple categories the macro-average method of calculating f-score is used [17].

Also 7-fold cross-validation algorithm was employed for testing. The F-score is computed for each fold and after that the average of all these F-scores is computed.

4.3 Experimental results

There are two main subjects for experiments.

The first one is related to the use of information from web pages found via hyperlinks. Here we compare 3 approaches: using only main page for classification, using main page and title + meta tags from other selected pages and using concatenation of main page with other selected pages.

Second subject concerns our approach of using and weighting markup tags in feature extraction. In order to build the baseline for this method we remove all markup tags after preprocessing (leaving the content of these tags) and apply pure TF-IDF algorithm [15] before the classification step.

Both subjects are tested in binary classification task and in categorization by product type task.

Binary *«e-commerce or not»* classification

In case of binary classification on *«e-commerce»* and *«non-e-commerce»* classes retrieving e-commerce web sites is more valuable for us than filtering not e-commerce. Thus we evaluate binary classifier with F-score of *«e-commerce»* class. The F-score results are listed in the table below:

Table 2. F-score of *«e-commerce»* class

Used web site information	pure TF-IDF	TF-IDF with Tag weighting
only main page	0.85	0.89
main page + meta and title from other pages	0.89	**0.94**
main page + whole other pages	0.86	0.92

While analyzing the results, we have found that considerable part (approximately 43% on average) of mistakes here is caused by customer-to-customer web sites which are not included in e-commerce web sites (in our classification), but which feature values are quite similar to B2B and B2C web sites.

Observed results show that the most efficient way of using information from other pages is to extract only meta tags and title. This approach ignores possible noise in other parts of additional pages and takes only the summary of these pages which is useful for detecting e-shops, while using whole content of the pages leads to some mistakes. For example, other pages may contain description of any e-shop or some other information which can increase the chances of false positive error.

As it can be seen from the results proposed, Tag Weighting method is more efficient than pure TF-IDF as it outperforms pure TF-IDF in all ways of extracting data. Obviously, most of e-commerce websites announce that they are e-shops in <title> and meta-information and Tag Weighting method weights give them maximum weight as these tags are unique. Also Tag Weighting is better at handling additional information from other pages as its F-score on «main page + whole other pages» is bigger than on «only main page» data. This is because Tag Weighting assigns small weight to some kinds of noisy information as it is usually located in frequently repeated tags.

E-commerce categorization

Table 5 lists the result of categorization expressed in macro-average F-score:

Table 3. macro-averaged F-score of e-commerce categorization by sold product type

Used web site information	pure TF-IDF	TF-IDF with Tag Weighting
only main page	0.67	0.72

main page + meta and title from other pages	0.74	0.79
main page + whole other pages	0.73	**0.81**

Again, Tag Weighting method performs better than pure TF-IDF. But in this case the most efficient was extracting the whole pages from hyperlinks found on main pages. This is due to the fact that the dataset for categorization contains only e-commerce web sites. This reduces the volume of noisy information on the pages of the website and thus decreases the chances of topic drift on different pages. Most highly specialized e-shops place hyperlinks to catalogue page or to some product categories pages or to certain product pages where they describe them thoroughly. This information is useful for classification by product type in most cases. Exceptions here are universal e-shops for which detailed descriptions of some goods may lead to misclassification. This can be seen on Table 6 which presents average F-score for each category for the case when Tag Weighting is used and main page is concatenated with whole other pages.

Table 4. F-score of e-commerce categorization by sold product type for each category

category id	category name	average F-score
0	Auto products	0.89
1	Medical goods	0.98
2	Health and beauty products	0.82
3	Appliances and electronics	0.79
4	Household goods	0.94
5	Furniture	0.92
6	Souvenirs and presents	0.69
7	Media (books, disks and concert tickets)	0.76
8	jewelry and clocks	0.73
9	Technical and industrial equipment	0.79
10	Food and kindred products	0.79
11	Pet supplies	0.85
12	Sport equipment and hobbies	0.73
13	Clothing and footwear	0.78
14	General stores	0.63

Significant number of misclassifications is connected with «Souvenirs and presents» and «Jewelry and clocks» categories because many web sites from these categories are very similar to each other. For instance, there are some web sites from «presents» category which sell clocks as an «expensive present». Also, there were many mistakes in «Sport equipment and hobbies» and «Clothing and footwear» categories because clothes and footwear are included in assortment of Sport equipment and hobbies» shops. Thus this is not a great surprise that the best classified categories are the least similar to others: «Auto product», «Medical goods», «Household goods» and «Furniture».

5 Conclusion

The main goal of the study was to understand which information can be useful in classifying web sites and how it can be used.

In order to check the hypothesis that the information from hyperlinks improves the quality of classification we have suggested the method for retrieving useful hyperlinks. We also compared some ways of using information from web pages found with these links. As a result, the experiments show that using the data from hyperlinks retrieved with our method increases the accuracy of classification. The experiments also revealed that it is preferably to exploit only meta tags and title from retrieved pages when diversity of data is high enough. Conversely, when the type of classifying data is more or less limited exploiting entire pages may improve the performance.

Another important idea was about exploiting the structure of a web page to enhance the classification. This paper introduces the approach of using weighted markup tags in feature extraction process and the idea of how to weight them. As illustrated by the experiment this approach is more efficient than feature extraction without considering the structure.

Proposed methods and approaches can be used not only in e-commerce classification but in any web classification task.

Also, some Russian e-commerce specific features were listed and explained.

The statement of the problem together with the interesting dataset gives a wide field of possible improvements and research:

(a)make the number of websites between categories more balanced and filter noisy web sites from datasets;
(b) test different machine learning algorithms and their ensembles on a current dataset;
(c)try to use hyperlinks which lead to another web site with filtering noisy hyperlinks and compare with the use of local hyperlinks only;
(d) develop some variations of Tag Weighting method;
(e)try to apply Fuzzy logic approach to categorization.

References

1. Grandon, E., Pearson, J.: Electronic commerce adoption: an empirical study of small and medium US businesses. Information & Management. 42, 197-216 (2004).
2. Yuan, S.: A personalized and integrative comparison-shopping engine and its applications. Decision Support Systems. 34, 139-156 (2003).
3. Wai Lam, Ruiz, M., Srinivasan, P.: Automatic text categorization and its application to text retrieval. IEEE Trans. Knowl. Data Eng. 11, 865-879 (1999).
4. Sebastiani, F.: Machine learning in automated text categorization. CSUR. 34, 1-47 (2002).
5. Apte C., Damerau, F., Weiss, S.: Automated learning of decision rules for text categorization. ACM Transactions on Information Systems. 12, 233-251 (1994).
6. Onan, A.: Classifier and feature set ensembles for web page classification. Journal of Information Science. (2015).
7. Qi, X., Davison, B.: Web page classification. CSUR. 41, 1-31 (2009).
8. Zengmin, G., Jianxia, D.: Research on web page classification-based core characteristics and web structure. International Journal of Wireless and Mobile Computing. 7, 253 (2014).
9. Chakrabarti, S., Dom, B., Indyk, P.: Enhanced hypertext categorization using hyperlinks. ACM SIGMOD Record. 27, 307-318 (1998).
10. Ghani, R., Slattery, S., Yang, Y.: Hypertext categorization using hyperlink patterns and meta data. ICML 01: Proceedings of the Eighteenth International Conference on Machine Learning. 178-185 (2001).
11. Oh, H., Myaeng, S., Lee, M.: A practical hypertext categorization method using links and incrementally available class information. SIGIR 00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 264-271 (2000).
12. Webpage Classification based on Compound of Using HTML Features & URL Features and Features of Sibling Pages. International Journal of Advancements in Computing Technology. 2, 36-46 (2010).
13. Vapnik, V., Cortez, C.: Support vector networks. Machine Learning. (1995).
14. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. 10th European Conference on Machine Learning. pp. 137-142. Springer Verlag, Heidelberg (1998).
15. Aizawa, A.: An information-theoretic perspective of tf-idf measures. Information Processing & Management. 39, 45-65 (2003).
16. Van Rijsbergen, C.: Information retrieval. Butterworths, London (1979).
17. Jackson, P., Moulinier, I.: Natural language processing for online applications. John Benjamins Pub., Amsterdam (2002).