Analysis and Evaluation of Web Pages Classification Techniques for Inappropriate Content Blocking

Igor Kotenko[1,2], Andrey Chechulin[1], Andrey Shorov[1,3], and Dmitry Komashinsky[4]

[1]St. Petersburg Institute for Informatics and Automation,
39, 14th Liniya, Saint-Petersburg, Russia
{ivkote,chechulin,shorov}@comsec.spb.ru
[2] St. Petersburg National Research University of Information Technologies,
Mechanics and Optics, 49, Kronverkskiy Prospekt, Saint-Petersburg, Russia
[3] Saint-Petersburg Electrotechnical University "LETI",
Professora Popova str. 5, Saint-Petersburg, Russia
[4] F-Secure Corporation, Tammasaarenkatu 7, PL 24, 00181 Helsinki, Finland
dmitriy.komashinskiy@f-secure.com

Abstract. The paper considers the problem of automated categorization of web sites for systems used to block web pages that contain inappropriate content. In the paper we applied the techniques of analysis of the text, html tags, URL addresses and other information using Machine Learning and Data Mining methods. Besides that, techniques of analysis of sites that provide information in different languages are suggested. Architecture and algorithms of the system for collecting, storing and analyzing data required for classification of sites are presented. Results of experiments on analysis of web sites' correspondence to different categories are given. Evaluation of the classification quality is performed. The classification system developed as a result of this work is implemented in F-Secure mass production systems performing analysis of web content.

Keywords: Classification of web pages, data mining, text analysis, HTML structure analysis, analysis of URL addresses, hierarchical classification.

1 Introduction

Nowadays, the Internet is one of the main ways of information obtaining. Lack of mechanisms that may classify information and control access to it in the Internet creates a problem of accessing unacceptable information by a certain circle of people. First of all, it concerns the need to restrict access to certain types of information according to age categories. In addition, limiting of automatic access to sites belonging to the higher risk categories (for example, "adult sites", "sites with unlicensed software", etc.) will increase security of users from malicious and unwanted software.

Contemporary web resources have a complex hierarchical structure and consist of multiple elements, including formatted text and graphical content, program code and links. This causes a number of problems inherent to the task of classifying web pages, with necessity to analyze colossal volumes of heterogeneous, often conflicting and

P. Perner (Ed.): ICDM 2014, LNAI 8557, pp. 39–54, 2014.

changing data. The main purpose of the presented work was to develop a classification system intended to block inappropriate content for deploying in F-Secure [8] software systems performing analysis of web content. The main theoretical contribution of the paper is to develop models, techniques and algorithms for web page classification based on data mining and machine learning. The developed models and techniques include generalizations of binary classifiers of web pages based on text information, the text content of individual structural elements (HTML-tags), their address information (Universal Resource Locator, URL), and combining of binary classifiers for obtaining an unequivocal classification result. Within the framework of the developed models and algorithms, there was created a system for classification of web pages based on text data, whose language is different from the language in which binary classifiers were trained (the training language is English, the languages of recognition are the main European languages, Chinese and others). Developed models, techniques and algorithms made it possible to build an automated system for classification of web sites, allowing to collect and process data required for training and testing of classifiers. A tool that uses trained classifier for rapid classification of web pages in different languages was developed.

The novelty of the research presented in the paper is in the integrated use of existing and modified approaches to classification of text and other information to determine the web page category. In addition, a new approach to the construction of dictionaries of text attributes was used at construction of new features that allowed enhancing the overall quality of the classification. The paper presents main elements of the research and development performed; its structure is organized as follows. *Second section* discusses the main results of relevant research. *Third section* provides a general description of the developed approach and the basic tasks solved during its development. Special attention is paid to the system characteristics used in the learning process. Essential aspects of procedures of receiving and processing data during the system training phase are presented in *fourth section.* Brief description of software implementation of the approach and the main results, causing selected decisions on the structure of decision-making procedures and organization of the system are described in *fifth and sixth sections* respectively. The main conclusions and further research directions are discussed in *seventh section*.

2 Related Work

The basic list of problems inherent to this subject area and approaches to their solution is presented in the review by Qi and Davison [24]. It summarizes the state of research at the end of the first decade of the XXI century. It discusses the possible variants for setting web page classification problem (binary, multiclass, hierarchical) and the main models for representation of entities used in the Internet as a set of interrelated elements. Differences of classification tasks for web pages and text are determined. On this basis, the list of features that are applicable to classification is specified. The basic applicable models for web page representation and special methods that use them are considered.

Let us consider the papers published after the appearance of this review. Shibu and others [27] considered the optimization issues of learning process in this class of systems due to the combined use of traditional procedures for selection of significant features with data Page Rank. Patil [22] investigated the applicability of Naive Bayes (NB) classifier for learning of web page classification systems within the individual groups of internal features of HTML documents. For classification of web pages Xu et al. [29] proposed the algorithm called Link Information Categorization (LIC), based on the k nearest neighbors (kNN) method. Its essence lies in the definition of the category of the classified web page based on analyzing links that other web pages make to this one. Calculation of relation level of web pages containing links to a particular category is done. The classification is performed by nine categories.

In general, it should be noted that most often used features that are applied for web page classification are extracted from the page text content. For instance, Dumais and Chen [6] separated concepts of web page text, header information and descriptive information service tag "meta". They implemented the Support Vector Machine (SVM) method. Lai and Wu [18] used two approaches to obtain necessary features for classification: meaningful term extraction and discriminative term selection. The main idea is to extract unknown words and phrases that belong to specific domain. Thus obtained terms are very specific to a particular domain. The classification was made according to five categories. Vector space model [26] was used for the classification.

Tsukada et al. [28] performed web pages classification by analyzing the nouns extracted from web pages. For removal of common words that are not related to a particular topic, and auxiliary words a stop-list was used. Choice of words needed for classifier training was implemented using the Apriori algorithm [1]. The method of Decision Trees (DT) was used for classification. Training was carried out on a set of objects belonging to five categories. In the work of Kwon et al. [16, 17] text features, extracted from different blocks of formatted text web pages, had three different levels of significance. Solution of the classification problem was achieved using the kNN.

Qi et al. [23] chose as the source of textual data not only the web page, but also its "neighbors", represented by the terms "parent", "descendant", "brother", etc. SVM and NB methods were used. An important result of this work is the demonstrated improvement of classification using data from neighboring web pages, which, however, leads to additional computational costs.

An alternative direction of research in the field of classification of web pages is the usage of text data extracted from their address (URL). Baykan et al. [2] performed search of tokens (i.e. search of individual words in the URL using the dictionary) and their decomposition into sequences of length from 4 to 8 characters to obtain mined features. Furthermore, the combination of all received n-grams, regardless of their length, was used. Classifiers models were based on SVM, NB and the maximum entropy method (MEM). Classification of web pages by URL was also realized in several papers of Kan et al. [11, 12]. The following features were implemented for training of classifiers: tokens, URL elements, structural patterns, association of abbreviations with full words combinations. SVM and MEM were used for classifier training.

Among relevant publications that explore aspects of SVM for web page classification we also mark works of Joachims [10], Dumais et al. [7] and Yang et al. [31]. Yang et al. also apply kNN, used by Calado et al. [3] and Lam et al. [19]. More information about applying NB can be found in the works by already mentioned Joachims [10], Dumais et al. [7], as well as Lewis [20], Chakrabarti et al. [4] and McCallum et al. [21].

3 General Approach to Websites Classification

3.1 Classification and Features Selection

For classifying the content of web pages we must take into consideration a number of features and particularities peculiar to the presentation of content in the Internet. Sources of information are web page addresses (URL) themselves, text content, text structure (HTML, Hyper Text Markup Language), structure of links, showing "location" in relation to other elements of web, and internal multimedia content [24].

For most ways of web page representation the source text information that can be extracted from its address (URL) and individual formatted elements of its content (HTML tags) as well as graphic information (such as images) are the most informative sources. In addition, historical information about the web page changes, when it is available, is of great importance.

Typically, in the first place the web content classification systems use the text displayed by web pages. The obtained results are applied to clarify the next steps of the decision process. The proposed approach to training is based on the application of machine learning for obtaining classifiers necessary to build a common classification system. Using the information obtained during raw data processing, the lists of major keywords for certain categories (based on the structural features) are formed. Direct classifiers training occurs on the base of text data obtained in the process of analysis of web pages belonging to predetermined topics. Preliminary data analysis showed that in some cases the conflict situations requiring resolution at the stage of preparing the input data for training are possible. An example of this conflict is the proximity of the text content of certain web pages, directly related to pharmacology, to the domain of drugs. Obviously, this may lead to errors such as "false positive" that stipulates wrong web page domain identification. The proposed solution to this problem is to break the thematic data sets into a number of categories, specifying one or another aspect of the topic. For example, a topic related to alcoholic beverages can be broken down into categories that reveal issues brewing, wine making and so on.

In the paper we consider the web pages belonging to a number of categories related to the following topics: "adult", "alcohol", "gambling", "tobacco", "dating", "drugs", "hate", "violence", "weapon", "religion", "occults". For descriptions of other topics and related categories we used generalized notion of "unknown".

3.2 Construction of Base and Combining Classifiers

We rely on the formal description of the problem of knowledge extraction from the Web [5] refined by Kleinberg et al. [13]. Based on the data available from the web

page address (URL) and the formatted text content (HTML), a three level decision scheme was developed.

Formal Statement of the Problem. It is given a set O of objects $\{o_1, o_2, \dots, o_n\}$ and the set C of labels of target classes $\{c_1, c_2, \dots, c_k\}$. Each object $o_i \in O$ can be represented by the corresponding element x_i of the set X of object descriptions, where x_i is a set $[x_{i,1}, x_{i,2}, \dots, x_{i,m}]$. Getting object descriptions is achieved by the conversion $f_{Map}: O \to X$, defined by applying a set of features $F = \{f_1, f_2, \dots, f_m\}$, each element of which is a transformation of the form $f: O \to V$ which puts for an element of the set of objects O an element of characteristic values V. There exists a dependence $f_{Target}: X \to C$, assigning to each object description x_i of the object o_i an element of class labels $c_j \in C$. It is necessary to construct an algorithm f_{Result}, approximating the target dependence f_{Target} on an accessible collection of objects $O_{Training} \in O$ and on the whole set of objects O.

Description of the Approach. Let h be the number of individual aspects, within the frames of which we study each element of the set O, and z is the number of features used to describe the aspects inherent in each of the designated categories, under a separate aspect. Then the total number m of elements of set F is equal to $k \times h \times z$, and the set of features is represented as $F = \bigcup_{l \in (1,h), j \in (1,z)} F_{l,j}$.

Prediction of the first level $p1_{i,l,j}$ about the belonging of description of the first level x_i of the object o_i to the category j under the aspect l is the result of applying the algorithm of the first level $a1_{l,j}: x_i \to P1_{i,l,j}$, $p1_{i,l,j} \in P1_{i,l,j}$, $P1_{i,l,j} = \{true, false\}$.

Description $d2_{i,l}$ of the second level of the object o_i within the aspect l is represented as a set $[p1_{i,l,1}, p1_{i,l,2}, \dots, p1_{i,l,k}]$ of predictions of the first level and characterizes the object belonging to each of the categories of the set C as a separate aspect.

Prediction $p2_{i,l}$ of the second level about belonging the description $d2_{i,l}$ to one of the target categories within the aspect l is the result of applying the algorithm of the second level $a2_l: d2_{i,l} \to C$.

Description $d3_i$ of the third level of the object o_i is represented as a set $[p2_{i,1}, p2_{i,2}, \dots, p2_{i,h}]$ of predictions of the second level and characterizes the object belonging to a category of the set C throughout the whole set A of aspects.

Prediction $p3_i$ of the third level on belonging the description $d3_i$ to one of the categories is the result of applying the algorithm of the third level $a3: d3_i \to C$.

Within the framework of the presented approach the development of an algorithm f_{Result} requires formation of a set of algorithms of the first, second and third levels.

Used Metrics. The degree of proximity f_{Result} and f_{Target} is assessed based on the F-measure metric that combines measures of accuracy and completeness of the solutions, which are in turn the characteristics determined by the errors' index of the first and second kind (false-positives and false-negative checks, respectively): $F_\beta = ((1 + \beta^2) \cdot tp)/((1 + \beta^2) \cdot tp + \beta^2 \cdot fn + fp)$, where tp is the number of correctly identified cases for a target category, fn is the number of false-negatives, fp is the number of false-positives, β is a coefficient of significance of false-negatives

comparing with false positives (within the frames of the results presented here, both types of errors are equal, in order to keep focusing on FP-proof solutions $\beta \ll 1$).

The developed decision making scheme is represented in Fig. 1. Elements of the first level are functional blocks oriented to certain categories. They use pre-trained classifiers giving judgments about whether a particular vector, characterizing the specific web page, refers to a certain category (thus the elements of the first level of the developed scheme solve the problem of binary classification). Each classifier is formed based on descriptions of objects in space of features inherent to the target domain (category). The need for this is confirmed by obvious use of specific terms and phrases within individual topics and categories. Thus, the key feature of the scheme at this level is the use of binary classifiers focused on individual categories and features inherent to them. The resulting solutions are hereinafter called as the first level predictions. Elements of the second level of the developed scheme use classifiers oriented for making decision about belonging the vector of descriptions of the given web page to one of the given domains (categories). Here the information obtained in the analysis of individual structural aspects of a web page (for example, data from web page address or its elements' formatting) is used. Descriptions going to the input of the aspect-oriented elements are formed by prediction values of the first level. The resulting solutions are hereinafter called as the second level predictions.

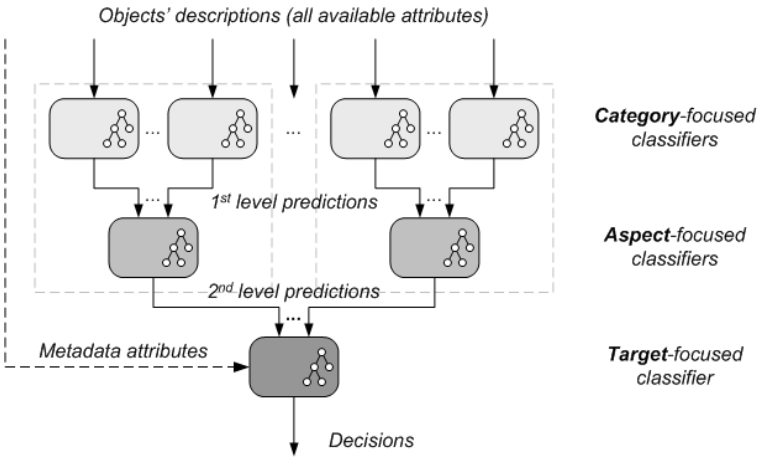

Fig. 1. Generalized representation of the developed decision making scheme

Predictions of the second level are used to generate descriptions of the analyzed web pages that are used for training and decision making by element of the third level, intended to form the final decisions.

Using this classification scheme was due to the decision of a number of research tasks, outlined in details in [14-15]. Firstly, we received a solution on the choice of the feature space used for training classifiers that are used by first level elements. It was shown that the use of relatively small sets of features relating to the target category is sufficient to create functional blocks of the scheme. Secondly, a set of experiments aimed at evaluation of the applicability of different groups of features used to

train the classifiers of the second and third levels of the scheme was performed. The result of this work was the decision to use at these levels only the solutions of the first level elements without involving the initial set of features.

To clarify the decision-making process, at the third level of the scheme in addition to the second level predictions, the information describing the context of obtaining web page data and general characteristics of its volume, origin, etc. may be used.

The proposed approach allows training a classifier consisting of atomic classifiers of different levels, specializing in working with certain types of data. Separate particularity of the approach is the ability to make quick update and supplement of the decision making scheme at entering new categories and domains or the emergence of new data sets useful for retraining already used decision making elements.

4 Obtaining and Processing of Data

Obtaining of raw data and their preparation for further analysis is one of the most important and effort consuming steps. This stage includes primary data loading, selection of necessary data, their retrieval from the web page and formation of combined features based on existing ones. This stage can be divided into the following sub-steps: (1) creating a list of web pages to build the training and test samples; (2) download and parsing of the contents of web pages; (3) extraction and formation of combined features by web pages content. As sources for the classification the following aspects of the data specifying a web page were selected: URL identifier, full text contained in the page and visible to the user, some text areas, marked with most thematically correlative tags and general statistics of tags.

In the research process the SVM, DT and NB classifiers were used. For each of the classifiers different combinations of classification parameter values were considered.

Let us consider different source data types in more details.

4.1 Processing and Analysis of Textual Information

The main source of data for the construction and performance of the trained model is the general web page text and the text contained in specific tags. To perform the analysis of the text by using the classifier, the text is presented in the form of keywords or phrases (wordforms) based on different methods of decomposition of the text.

There are many ways to split the text into individual wordforms. During performing this research the analysis of the following methods was done: (1) separating at appearance of non-alphanumeric characters, (2) separation into clauses on the basis of punctuation analysis, and (3) division into tokens (based on the dictionary of Rapid Miner [25]).

We also analyzed the following text preprocessing techniques: (1) stemming (process of finding a basis for a given word of the original word); (2) replacement of individual words with their hyponyms (based on the Oxford Dictionary); (3) replacement of individual words with their hyperonyms (based on the Oxford Dictionary).

In addition, we built a list of prepositions, conjunctions, pronouns, etc. that do not have any independent meaning, and in the case of the analysis of individual words the elements of this list were excluded from processing.

Let us consider in more detail the technique of building the vocabulary of basic wordforms for further training of the classifier. This technique is based on the analysis of the frequency characteristics. It involves the following steps: (1) Preparation of text content for all web pages (here we can use both the general text and the text stored in individual tags). (2) Exception of stop words from all texts (i.e. words that do not carry the semantic load). Efficiency of the inclusion of this step was checked in experiments. (3) Processing of all texts by stemmer (i.e. separation of the part that is the same for all the grammatical forms of a word by cutting off endings and suffixes). Efficiency of the inclusion of this step was checked in experiments. (4) Creation of the lists of keywords for each category with given values of TF (term frequency), where TF is the value calculated as the ratio of the number of occurrences of a word to the total amount of words in all web pages of the category, defining the importance of the words within a category. (5) Calculation of IDF (inverse document frequency) that is inversion of frequency with which a word occurs in the selected categories. Using the IDF allows to create the word vector of high importance within one category and rarely found in other categories.

When forming the dictionary we also used an original technique for calculation of TF and IDF. TF is normally calculated as the ratio of the number of occurrences of a word to the total number of words in the document. In our research, a category was used in the role of a document, and the number of occurrences was determined as the number of web pages containing this word. At the same time we investigated various indicators characterizing the presence of words in the text (for example, a single occurrence of the word in the text, consisting of 100 thousand words, does not mean that the text is relevant to the category to which this word belongs). The same approach was used for IDF, i.e. the word having low value of TF for some category may be considered as absent in this category. In experiments we checked different boundary parameters determining occurrence of a word in a text or a category, and found their optimal values.

After completion of the previous stage, for each web page a table is filled in which a value from the set {0,1} corresponds to each wordform. This value determines whether this web page specific contains a wordform belonging to a category.

In the next step the training of atomic classifiers for particular categories is performed. Each classifier receives on the input a limited number of wordforms that are the most relevant to one category or another. Thus, for training the classifier, the following source data for each web page were used: (1) the number of occurrences of each wordform from the dictionary in the text of the web page (in the experiments we considered different sizes of dictionaries), (2) the category of the web page.

After comparing the quality of text classification using various techniques we chose the technique based on decision trees, as it showed high level of precision along with adequate recall, accuracy and F-measure.

4.2 Processing and Analysis of URL Addresses of Web Pages

To analyze the URL addresses of web pages we used a technique similar to the approach used in the analysis of textual information. However, given the particularities of URL, the selection of individual elements of the text was performed with a method of finding n-grams. After that for these n-grams we used the same steps as for the wordforms in text analysis (TF-IDF).

Let us consider in more details the technique for building the dictionary of n-grams for further training of the classifier. This technique comprises the following steps: (1) collection of URL addresses of all web pages belonging to the same category in one text object (thus, for each category a separate text object was formed); (2) formation of text tokens based on special characters that are contained in URL addresses; (3) removing of stop words and applying algorithms of stemming to obtained text tokens; (4) generation of n-grams for processed lists of tokens. Thus, for training a classifier the following input data for each web page were used: presence of each n-gram from dictionary in URL address of the web page; integer attribute that characterizes the total number of n-grams presented in the URL address of the web page and specific to a particular category; web page category.

4.3 Processing and Analysis of Information on Structure of Web Pages

To identify such categories as news, forums, blogs, etc. an approach was used, based on the analysis of the HTML structure of web pages. An important peculiarity of these categories is heterogeneous text content and attributes contained in the structure of tags (number of references, tables, headers, etc.). To select tags that allow to identify web pages of certain categories, the information weight of each tag was calculated, and the sample of tags, based on a predetermined boundary, was produced. For training of the classifier the following input data for each web page were used: (1) the number of occurrences of each of the selected tags in the HTML code of the web page, (2) the length of the text in different tags, and (3) the category of the web page.

4.4 Classification of Sites in Languages Other Than English

To classify sites in foreign languages it is suggested to use translation of text content of web pages from the original language into the language that was used for training of classifiers. This approach allows to classify web pages in any language supported by the system of automatic translation, using models of classifiers trained on English web pages. There is no need to prepare additional classifiers to categorize web pages for each specific language. This eliminates the need for preparation of additional training data sets and further maintenance of them in up to date state. In its turn, the use of classifiers focused on a specific language allows to use different grammatical structures for training that can improve the classification quality.

4.5 Other Sources of Information about Web Pages

In the research we also considered other sources of data based on which it is possible to get information characterizing the category to which the web page belongs.

For example, the following sources were analyzed: (1) data about web pages that are referenced by the analyzed web page, or which in turn refer to it; (2) information from WhoIs servers; (3) existing lists of web pages containing those labels (for example, lists used in parental control systems); (4) defined category of the web page for a certain period of time (it allows to take into account changes in the web site content).

The mentioned sources currently do not participate in the process of classification for the following reasons: (1) Analysis of data on the links between sites is a perspective direction, but it requires a large and coherent training sample of web pages, the collection of which is a separate challenge. (2) When analyzing the responses of WhoIs servers we experienced difficulties caused by the complexity of messages analysis from different servers; it is a consequence of the lack of a common format of such answers. (3) Using existing lists of classified sites may have a place in the final classification scheme as an expert opinion, allowing to evaluate the result of the system, and as one of the sources of construction of provisional estimation of the site belonging to one category or another. In connection with the low quality of the original data in public databases, these lists were used only as part of the training and testing samples for the analysis of the quality of other classification techniques. (4) History of previous classifications should be accumulated over a sufficiently long period of time, so the research on evaluation of this information source continues.

5 Implementation

For downloading data from the sites, training classifiers and their testing we developed a software tool, called Web Classification Manager. The tool is developed in Java programming language and uses the following software libraries: Jsoup HTML parser 1.7.1 [9] and Rapid Miner 5.2 [25].

The software tool operates in the following three main modes:

1. Source data preparation mode. This mode is used for the preparation of the training and testing databases that do not depend on changes in the Internet. At the same time according to the prepared list of URL addresses of already classified web pages their HTML representation is loaded, from which, in turn, attributes for training of classifiers are selected. The list of classified web pages serves as the input data.

2. Training mode. This mode is designed for automated preparation of the trained model of the web page classifier. The features of classified web pages act as the input data. The result is a trained classifier model.

3. Testing mode. It allows assessing the quality of the trained model. Indicators for which a classifier is being evaluated are described in details above. The features of classified web pages and the trained classifier model serve as the input data. The result is a set of indicators characterizing the efficiency of the trained classifier model.

The tool provides the ability to perform in automated mode a wide variety of operations, and provides a “one-touch” tool for training and testing of the classifier models. During testing of the tool, we revealed that when the training set of web pages consists of 100 000 URL addresses preparation time and testing of the model takes about 70 hours (including 10-15 minutes of operator's work). Using the tool, we

prepared: (1) the training and testing samples of web pages, as well as (2) the trained basic and combined classifiers based on the analysis of different kinds of information.

6 Experiments and Collected Statistics

6.1 Preparation of Training and Testing Samples

This stage can be divided into the following sub-steps: (1) creation of the list of categories of web pages; (2) preparation of the input lists of URL addresses of web pages; (3) loading the web content to the internal storage; (4) data pre-processing and extraction of features, which will be used to train classifier models.

Creation the list of categories of web pages. Quality of trained classifiers largely depends on the chosen categories of classification. For example, in some taxonomies, categorization of "hate" and "cruelty" are separated, however, there are many sites that cannot be clearly attributed to one or another category. Similar problems are "hacking", etc. Providing the list of categories it is necessary to maintain balance of detail and precision. Presence in the same list of categories such as "computers" and "software", "sport" and "fitness", etc. can cause reduction of classification quality.

Preparation and analysis of the initial lists of URL addresses of web pages. In this sub-step, the problems can be associated with using various sources of classified web pages for different categories. This occurs because the boundary between the categories is often subjective, leading to problems when training the classifier.

Loading the web content to the internal storage. Here the additional complexity is given by web pages with dynamic content (for example, web pages with Javascript). The content of such pages may depend on many factors and, as a consequence, the loaded page may not always match what the user can see through the browser.

Data pre-processing. The main objective here is to choose the most informative features that characterize the categories by which classification will be carried out. For example, category "news" is extremely difficult to determine on textual content, but the web pages of this category often have similar structure tags. Thus, the list of selected indicators directly depends on the selected categories.

As the *input data source* the Open Directory Project (ODP), also known as DMOZ was chosen. After several experiments it was found that it contains a lot of misclassified sites that decrease the further classification precision. On the last stages of experiments the manually tested site's list provided by F-Secure company was used.

6.2 Classification by Textual Information

To assess the quality of work of various classification techniques the cross-validation method was used (see Table 1).

Table 1. Comparison of quality of work of classifiers based on SVM, DT and NB

	Precision, %	Recall, %	Accuracy, %	F-measure, %
SVM	83.44	15,83	16,44	16,95
DT	82,34	17,66	17,72	20,75
NB	74,68	39,96	39,72	43,54

After comparison of quality of classification of text information using a variety of machine learning methods, the DT method was selected, as it showed a quite high level of precision with adequate results in recall, accuracy and F-measure.

Based on the results of the experiments carried out we can make the following conclusions: (1) On experiments, dealing with finding the optimal size of feature dictionary, we can conclude that the expansion of the dictionary improves the accuracy of classification only up to some limit, and after that information saturation leads to the fact that accuracy has been increasing slightly. (2) Each classifier determines with maximal accuracy different categories, hence the use of combination of individual classifiers of different types in different categories will lead to a significant increase in accuracy. (3) Introduction of a new category "unknown" and attributing to it of all web pages that cannot be exactly attributed, improved the accuracy of combined classifiers. Fig. 2 shows the analysis of different approaches to the choice of wordforms for constructing a dictionary (division into words based on analysis of non-alphabetic symbols, separation into sentences, separation into tokens, replacing words by hyponyms, replacing words by hyperonyms). The figure shows that the best accuracy was demonstrated by the method of partitioning the text on the basis of tokens.

6.3 Classification Based on URL Addresses

Experiments have shown that each of the categories has its specific well-defined set of n-grams, appearance of which in URL address is more likely.

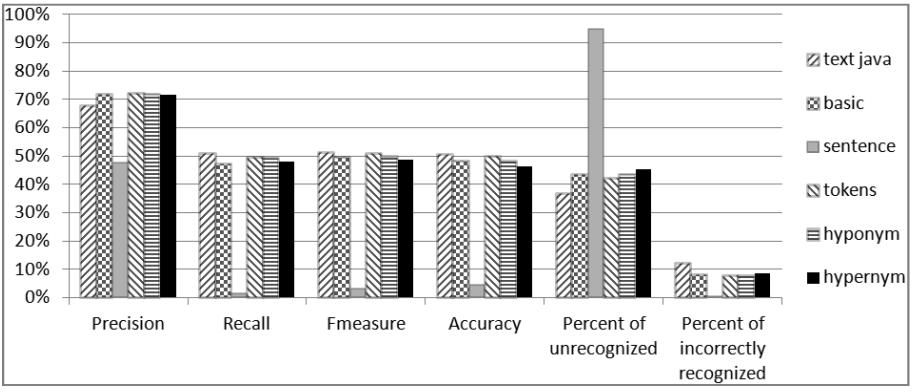

Fig. 2. Comparison of different methods of text splitting

However, in the general case, such n-grams are not necessarily present in the URL addresses of the corresponding categories. This explains the relatively high values of the precision of the target class at low values of the recall. This means that if a separate classifier trained on the attributes retrieved from the URL address, informs that the object belongs to its target category, then it can be trusted. However, due to the specificity of these data, there will be a set of objects not related to any category. The experimental results also showed that the technique of classification based on n-grams can be used in a complex solution that is based on multi-expert paradigm where each expert is responsible for decision making on attributing the classified URL address to the target category. From such a decision one may expect sufficiently high values of the precision, but it will be efficient for the 25-40% of the targets. Possible ways to enhance this approach is to use a more balanced and filled list of features and data.

6.4 Combined Classification

In general, the classifier based on the analysis of textual information shows the best results (high precision and recall) and the classifier based on URL addresses gives greater precision compared with the analysis of the structure of HTML. However, the combined use of individual classifiers based on these aspects of data that describe the object, gives significant improvement in the classification results.

Fig. 3 shows the results of the classification of the test data set using models trained on the text, models trained on the URL, the combined model trained on the text and tags, and the combined model trained on the text, tags and URL. This figure shows the values of indicators such as precision, recall, F-measure, accuracy, the percentage of "unknown" and misclassified web pages. The experimental results confirm the increase in classification performance when using the combined classification approach.

6.5 Classification of Web Pages in Languages Other Than English

To classify web pages in languages other than English, the first phases of preparation of input data of the testing sample were modified.

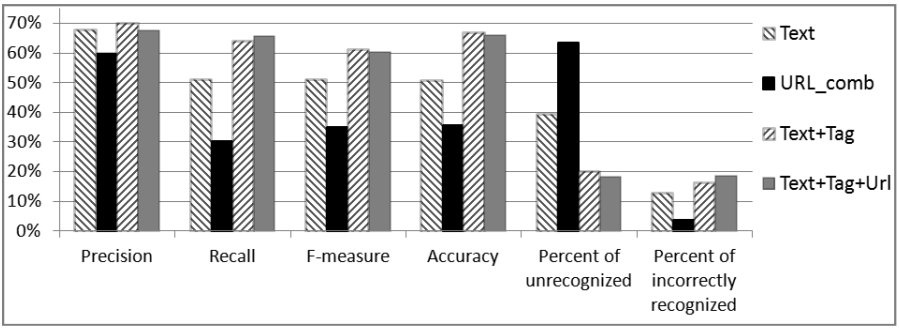

Fig. 3. Results of classification of testing data set

After loading the text content from web pages, the translation into English (which was used for the training of classifiers) was performed. Machine translation was performed using Yandex.Translate system [30], and the language of the original web page was determined automatically. Further stages of the web pages classification are similar to those performed in the classification of sites in English.

To verify the proposed approach the testing samples were formed for web pages belonging to the category "adult". For sites on the English, French and German the results are 94, 91 and 96%, respectively. These results show high classification quality of the "adult" category in German and French. However, it should be noted that the testing sample contains only the sites of these categories. In the case of inclusion in the sample of web pages belonging to other categories, F-measure is slightly reduced, as the classifier is likely to include web pages of other categories into the category of "adult" (type II errors).

It should be also noted that the quality of classification of web pages in German was slightly better than a similar indicator for English. This is mainly due the fact that some words having different connotations are translated as a single word, which reduces the diversity of wordforms, and as a result, can lead to better classification.

6.6 Comparison with Existing Analogues

For making comparison within the limits defined above, there were considered indicators of accuracy given in the papers of Chakrabarti et al. [4], Qi et al. [23], Calado et al. [3], Patil et al. [22] (75, 91, 81 and 89%, respectively) that have values that do not contradict principally to the values that we obtained (average accuracy across all available categories is about 67%). However, it should be noted that the result obtained of this paper was averaged. It does not display the accuracy achieved for certain categories (in the experiments we considered more than 20 categories, and, for example, for the categories "cigarettes", "dating", "adult", "casino" and "marijuana" the accuracy ranged from 87% up to 96%, but at the same time, the categories "cults", "occultism", "racism" and "religion" showed less than 50% accuracy), and it also ignores the importance of the additional "compensatory" mechanism counteracting the problem of false positives by using "unknown" as a category used to classify the "default" object (in this case the percentage of correctly classified sites was 67%, while the percentage of "unknown" sites was 15%, and the error rate was only 18%).

7 Conclusion

The paper proposed an approach to classify web information by applying Machine Learning and Data Mining techniques. The approach automates preparing input data and trained models. Architecture and algorithms of the system for collecting, storing and analyzing the data needed to classify web sites into certain categories were proposed. The developed architecture of combining base classifiers into a general scheme uses advantages of individual classifiers and therefore neutralizes their limitations. The software system to automate the classification of web pages was implemented.

Experiments to identify the main problems in the construction of web page classifiers were carried out. During experiments a lot of time was spent for understanding why certain classifier's decisions are wrong or good, testing different hypothesis and checking solutions. As a result, experiments showed high FP-proof classification accuracy for certain categories. This confirms feasibility of using the technology in systems of blocking websites with inappropriate content.

Further work on the development of the approach includes research areas aimed at improving the quality of decisions taken by the developed system. Very important goal of future work is taking context into account in text analysis.

Acknowledgements. This research is being supported by TEKES as part of the Data to Intelligence program of DIGILE (Finnish Strategic Centre for Science, Technology and Innovation in the field of ICT and digital business), the grants of the Russian Foundation of Basic Research (13-01-00843, 13-07-13159, 14-07-00697, 14-07-00417), the Program of fundamental research of the Department for Nanotechnologies and Informational Technologies of the Russian Academy of Sciences, the state project "Organization of scientific research" of the main part of the state plan of the Board of Education of Russia, and the SPbNRU ITMO project.

References

1. Agrwal, R., Srikant, R.: First algorithms for mining association rules. In: Proc. of the 20th Very Large Data Bases Conference, pp. 487–499 (1994)
2. Baykan, E., Henzinger, M., Marian, L., Weber, I.: Purely URL-based topic classification. In: Proc. of the WWW 2009, New York, USA, pp. 1109–1110 (2009)
3. Calado, P., Cristo, M., Moura, E., Ziviani, N., Ribeiro-Neto, B., Goncalves, M.A.: Combining link-based and content-based methods for web document classification. In: Proc. of the CIKM 2003, New York, USA, pp. 394–401 (2003)
4. Chakrabarti, S., Dom, B., Agrawal, R., Raghavan, P.: Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. The Intern. Journ. on Very Large Data Bases 7(3), 163–178 (1998)
5. Cooley, R., Mobasher, B., Srivastava, J.: Web mining: Information and pattern discovery on the world wide web. In: Proc. of the ICTAI 1997, pp. 558–567 (1997)
6. Dumais, S., Chen, H.: Hierarchical classification of Web content. In: Proc. of the SIGIR 2000, pp. 256–263. ACM, New York (2000)
7. Dumais, S.T., Platt, J., Heckermann, D., Sahami, M.: Inductive learning algorithms and representations for text categorization. In: Proc. of the CIKM 1998, pp. 148–155 (1998)
8. F-Secure company, `http://www.f-secure.com/`
9. Java HTML Parser, `http://jsoup.org/`
10. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
11. Kan, M.Y., Thi, H.O.N.: Fast webpage classification using url features. In: Proc. of the CIKM 2005, New York, USA, pp. 325–326 (2005)
12. Kan, M.Y.: Web page classification without the web page. In: Proc. of the WWW Alt. 2004, New York, USA, pp. 262–263 (2004)

13. Kleinberg, J.M., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.S.: The Web as a Graph: Measurements, Models, and Methods. In: Asano, T., Imai, H., Lee, D.T., Nakano, S., Tokuyama, T. (eds.) COCOON 1999. LNCS, vol. 1627, pp. 1–17. Springer, Heidelberg (1999)
14. Komashinskiy, D.V., Kotenko, I.V., Chechulin, A.A.: Categorization of web sites for inadmissible web pages blocking. High Availability Systems (2), 102–106 (2011)
15. Kotenko, I.V., Chechulin, A.A., Shorov, A.V., Komashinkiy, D.V.: Automatic system for categorization of websites for blocking web pages with inappropriate. High Availability Systems (3), 119–127 (2013)
16. Kwon, O.W., Lee, J.H.: Text categorization based on k-nearest neighbor approach for web site classification. Information Processing and Management: an International Journal 29(1), 25–44 (2003)
17. Kwon, O.W., Lee, J.H.: Web page classification based on k-nearest neighbor approach. In: Proc. of the IRAL 2000, New York, USA, pp. 9–15 (2000)
18. Lai, Y.S., Wu, C.H.: Meaningful term extraction and discriminative term selection in text categorization via unknown-word methodology. ACM Transactions on Asian Language Information Processing (TALIP) 1(1), 34–64 (2002)
19. Lam, W., Ho, C.Y.: Using a generalized instance set for automatic text categorization. In: Proc. of the SIGIR 1998, Melbourne, Australia, pp. 81–89 (1998)
20. Lewis, D.D.: An evaluation of phrasal and clustered representations on a text categorization task. In: Proc. of the SIGIR 1992, Copenhagen, Denmark, pp. 37–50 (1992)
21. McCallum, A., Nigam, K.: A comparison of event models for naive Bayes text classification. In: Proc. of the AAAI/ICML 1998, pp. 41–48. AAAI Press (1998)
22. Patil, A., Pawar, B.: Automated Classification of Web Sites using Naive Bayessian Algorithm. In: Proc. of the IMECS 2012, vol. 1, p. 466 (2012)
23. Qi, X., Davison, B.D.: Knowing a Web Page by the Company It Keeps. In: Proc. of the CIKM 2006, pp. 228–237 (2006)
24. Qi, X., Davison, B.D.: Web Page Classification: Features and algorithms. ACM Computing Surveys (CSUR) 41(2), article No.12 (2009)
25. RapidMiner, `http://rapid-i.com/content/view/181/190/`
26. Schauble, P.: Multimedia Information Retrieval: Content-Based Information Retrieval from Large Text and Audio Databases. The Springer International Series in Engineering and Computer Science, pp. 49–59. Kluwer Academic Publishers, Norwell (1997)
27. Shibu, S., Vishwakarma, A., Bhargava, N.: A combination approach for Web Page Classification using Page Rank and Feature Selection Technique. International Journal of Computer Theory and Engineering 2(6), 897–900 (2010)
28. Tsukada, M., Washio, T., Motoda, H.: Automatic Web-Page Classification by Using Machine Learning Methods. In: Zhong, N., Yao, Y., Ohsuga, S., Liu, J. (eds.) WI 2001. LNCS (LNAI), vol. 2198, pp. 303–313. Springer, Heidelberg (2001)
29. Xu, Z., Yan, F., Qin, J., Zhu, H.: A Web Page Classification Algorithm Based on Link Information. In: Proc. of the DCABES 2011, pp. 82–86. IEEE Computer Society (2011)
30. Yandex. Translate API: `http://api.yandex.com/translate/`
31. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proc. of the SIGIR 1999, Berkeley, CA, pp. 42–49 (1999)