Detecting Content Spam on the Web through Text Diversity Analysis

© Anton Pavlov

M.V. Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics
pavvloff@yandex.ru

© Boris Dobrov

M.V. Lomonosov Moscow State University, Research Computer Center
dobroff@mail.cir.ru

Abstract

Web spam is considered to be one of the greatest threats to modern search engines. Spammers use a wide range of content generation techniques known as content spam to fill search results with low quality pages. We argue that content spam must be tackled using a wide range of content quality features. In this paper we propose a set of content diversity features based on frequency rank distributions for terms and topics. We combine them with a wide range of other content features to produce a content spam classifier that outperforms existing results.

1 Introduction

Web spam or spamdexing is defined as "any deliberate action that is meant to trigger an unjustifiably favorable relevance or importance for some Web page, considering the page's true value" [15]. Studies show that at least 20 percent of hosts on the Web are spam [7]. Web spam is widely acknowledged as one of the most important challenges to web search engines [16].

There is a wide range of spamming techniques usually aimed at different algorithms used in search engines. This article is dedicated to content spam detection algorithms. Content spamming or term spamming refers to "techniques that tailor the contents of text fields in order to make spam pages relevant for some queries" [15]. We argue that content spam can be detected using a combination of text quality features that cover multiple characteristics of natural texts. In this work we introduce several novel features based on frequency rank distributions for terms and topics that substantially improve content spam classification.

In Section 2 we provide basic assumptions behind our research. In Section 3 we describe the content spam detection framework. Section 4 contains evaluation results. Section 5 is dedicated to future work and conclusions.

Proceedings of the Spring Young Researcher's Colloquium On Database and Information Systems SYRCoDIS, Moscow, Russia, 2011

1.1 Related Work

Many spam detection techniques have been proposed in recent years during the Web Spam Challenge [20]. Some content features we used were proposed by Ntoulas et. al. [17]. This work showed that compressibility of text and some HTML-related characteristics distinguish content spam from normal pages. A large amount of linguistic features were explored in a work by Piskorski et. al. [19]. Latent Dirichlet Allocation [5] is known to perform well in text classification tasks. Biro et al. did a lot of research on modifying the LDA model to suit Web spam detection. They developed the multi-corpus LDA [3] and linked LDA [4] models. The former builds separate LDA models for spam and ham and uses topic weights as classification features. The later incorporates the link data into LDA model to spam classification.

Web spam is also aimed at web graph features used by search engines so many researchers focused on detecting link spam. Techniques like TrustRank [14] minimize the impact spam pages in ranking. Much attention has been focused on fighting link farms – web graph structures designed to accumulate PageRank and affect other pages rankings [21]. Finally more and more researchers combine the link and content data to improve classification results [1, 4]. In this work we didn't use any link spam detection techniques as we focused on content spam.

Fetterly et. al. proposed using duplicates analysis to detect web spam [10]. They measured phrase-level duplication of content across the web and found that spam tends to have greater number of popular shingles per document.

2 Understanding Content Spam

We believe that tackling Web spam is impossible without understanding how it works. Content spamming is aimed at text relevance algorithms, such as BM25 and tf.idf [15]. These algorithms are particularly vulnerable to content spam as there is a strong

correlation between document relevance and amount of query terms found in the text.

Content spam is often used in doorways – pages and sites designed specifically to attract and redirect traffic. Doorways are only efficient if they reach the top of search results. Spammers prefer to generate thousands of doorway pages, each optimized for a specific query, to maximize amount of traffic collected.

This leads to several requirements that content spam must satisfy to be efficient:

- It must be generated in thousands of pages;
- Each page must maximize text relevance for some search query.

Thus spammers have very little options of generating content for their doorways:

- They may generate content automatically;
- They may duplicate texts from other web sites;
- Or they may use a combination of both techniques.

Automatic text generation is a difficult task that does not have a satisfactory solution yet. Natural texts have multiple levels of consistency that are extremely hard to emulate all at once. In text generation tasks such as automatic document summarization researchers distinguish multiple qualities of natural texts. Experiments show that even specialized text generation algorithms score low in most of these measures [9].

The levels of consistency include local coherence, style and authorship consistency, topical consistency, logical structure of the document etc. In this setting the uniqueness of text is just another type of constraint that is inherent for natural texts. Our approach is based on controlling as much natural text constraints as possible, making it harder for spammers to conceal low quality content.

There is a wide range of text generation techniques that generate locally coherent yet unreadable texts. Techniques like Markovian text generators are often used by web spammers to generate unique texts in great numbers. We were especially interested in designing text quality analyzer that would detect such advanced types of web spam.

3 Content Spam Detection Framework

Our work was based on assumption that spammers cannot emulate all aspects of natural texts. Our goal was to address as many domains of consistency as possible, by using various features. We measured multiple aspects of text quality and used supervised learning to combine them into content spam classifier. Despite a popular trend of combining link and content detection methods we focused solely on content analysis.

The basic natural language characteristics such as readability and POS ratios are overviewed in Section 3.1. The novel part of our spam detection framework is a set of text diversity features. We designed a range diversity features based on frequency rank distributions for different aspects of text diversity. The description and analysis of these features are provided in Section 3.2. Topical classification and topical diversity features based on LDA statistical model are presented in Section 3.3.

All statistics on described features were collected on WEBSPAM-UK2007 dataset [22]. The spam prevalence histograms provided in this section were generated on the set of 3995 labeled hosts from the training part of the dataset.

3.1 Statistical Features

The benefit of using wide range of linguistic features has been shown before by Piskorski et. al [19]. These features are commonly used in stylometry and authorship identification. We used POS tagger to tag every word in the dataset. We also substantially elaborated linguistic features by implementing a set of style-related diversity features that are described in Section 3.2.

In order to extract maximum information from POS tagging we calculated ratios of different parts of speech in words and ratios of different grammatical categories:

- POS ratios:
  - Adjectives;
  - Nouns;
  - Pronouns;
  - Verbs;
  - Numerals;
  - Particles;
  - Conjunctions;
  - Articles;
- Grammatical categories:
  - Number;
  - Tense;
  - Aspect;
  - Mood.

Combinations of different parts of speech and categories resulted in 82 distinct grammatical forms. We calculated the ratio of each grammatical form:

$$Ratio(form) = \frac{\#\ form\_occurences}{\#\ words}.$$

We also measured ratios of grammatical categories for specific parts of speech, e.g. ratio of verbs in past tense compared to all verbs:

$$Ratio_{verbs}(past\_tense) = \frac{\#\ verb\_in\_past\_tense}{\#\ verbs}.$$

As a result, we used a total of 145 POS-related statistical features.

Another domain we took features from was text readability research. Readability metrics were developed for military and educational purposes to measure how hard the text is to understand. Such features are helpful as automatically generated texts are usually unreadable. Some readability features have already been investigated by Ntoulas et. al. [17]. We implemented a set of readability features:

- Average word length;
- Average sentence length;

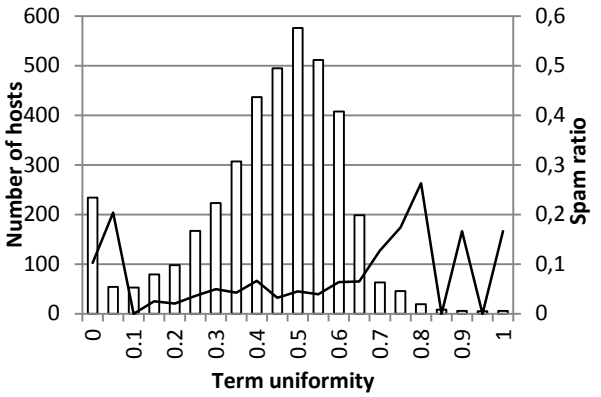

Figure 1. Prevalence of spam relative to term uniformity

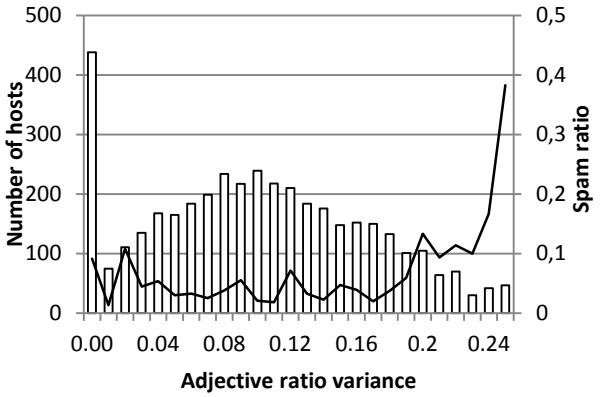

Figure 2. Prevalence of spam relative to adjective ratio variance across sentences

- Average number of punctuation symbols per sentence;
- Ratio of words longer than 7 symbols;
- Ratio of words shorter than 3 symbols;
- Maximum sentence length;
- Minimum sentence length.

The set of 152 statistical features described above allows detecting simple anomalies in text, such as query dumping, but it is still inadequate to fight advanced types of spam.

3.2 Text Diversity Features

Many researchers noticed that entropy and compressibility distinguishes content spam from normal texts [17]. We argue that this trait stems from auto generated nature of content spam. Currently no text generation algorithm can repeat the variety of natural language.

Some diversity-related features are easily faked by spammers. It is not uncommon for content spammers to use garbage text to decrease compressibility of texts in attempt to foil spam detection algorithms. To overcome these limitations we propose measuring variety of content in multiple aspects.

3.2.1 Character-Level Diversity

Compressibility is a well-known text variety feature. This characteristic has been used in both e-mail [6] and web spam detection [17]. Some content spamming techniques such as keyword stuffing produce texts with large number of repetitions. We use gzip and bz2 compression algorithms to measure compressibility of a document.

3.2.2 Term-Level Diversity

Compressibility is known to work well, when repeated keywords are located nearby in text. Spammers often dilute normal texts with keywords, thus making them harder to detect. Such subtle statistical violations can be detected by analyzing word frequency distributions.

Words in natural texts are known to obey power law frequency distributions. The most notable is Zipf law [23] that states that the frequency of any term is inversely proportional to its rank. Given a word *w*, with a frequency rank of *rank(w)*, its frequency may be estimated using the following formula:

$$freq(w) \approx \frac{const}{rank(w)^s}.$$

Parameter *s* characterizes variety of words in the given corpus of texts. We will refer to this value as to uniformity of terms. Greater uniformity leads to greater frequency of the most probable words, and lower frequencies of other words. The easiest way to calculate uniformity for a document is to convert the Zipf law to logarithmic scale:

$$\log(freq(w)) \approx s \log(rank(w)) + const.$$

Using this equation uniformity can be estimated using linear least squares. Let *n* be the number of different words in text, then:

$$f_w = \log(freq(w));$$

$$r_w = \log(rank(w));$$

$$s = -\frac{n\sum_w r_w f_w - \sum_w r_w \sum_w f_w}{n\sum_w (r_w)^2 - \left(\sum_w r_w\right)^2}. \qquad (*)$$

We estimated terms uniformity to detect texts that contain multiple repeating keywords. In order to reduce the effect of stopwords we also calculated uniformity for nouns.

We also used a simpler approximation of term-level diversity by calculating the average number of terms that are repeated in neighbor sentences.

The prevalence of spam relative to term uniformity is shown in Figure 1. In this figure the horizontal axis corresponds to different levels of term uniformity. The white bars correspond to number of hosts from the WEBSPAM-UK2007 training set with a given level of term uniformity and the black line corresponds to ratio of spam among those hosts. The figure shows that content spam tends to have greater uniformity, as spammers often repeat search keywords.

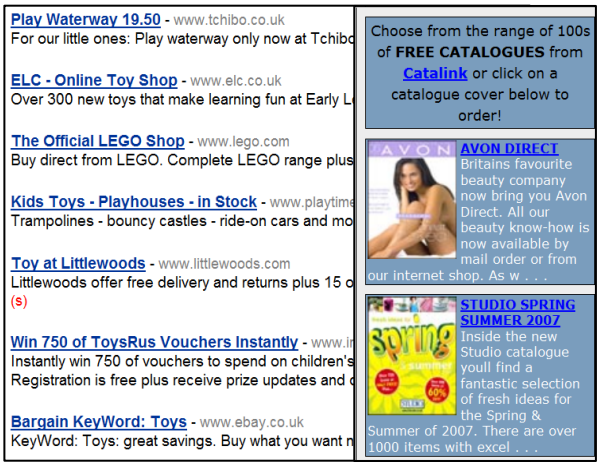

Figure 3a. Sample spam page with low topical uniformity (http://www.harrogate-toy-xmas-fair.co.uk/). The page consists of excerpts from different sources.

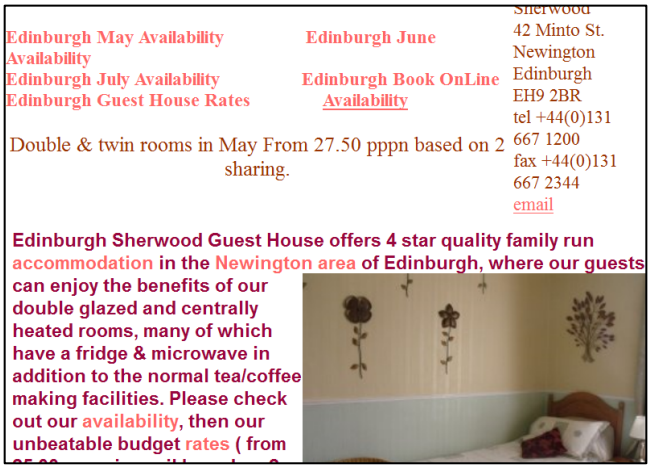

Figure 3b. Sample spam page with high topical uniformity (http://www.sherwoodguesthouseedinburgh.co.uk/). Notice keywords in the top and side of the page and highlighted keywords in text.

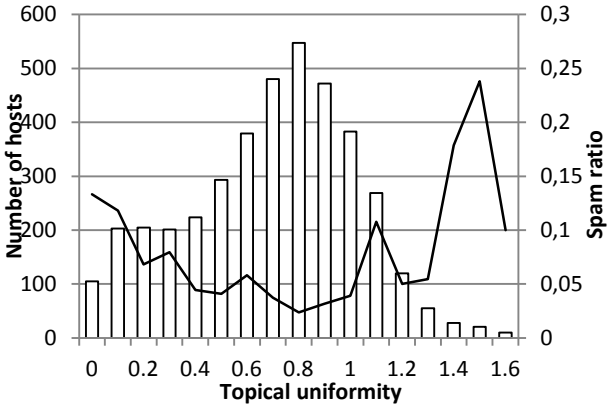

Figure 4. Prevalence of spam relative to topical uniformity

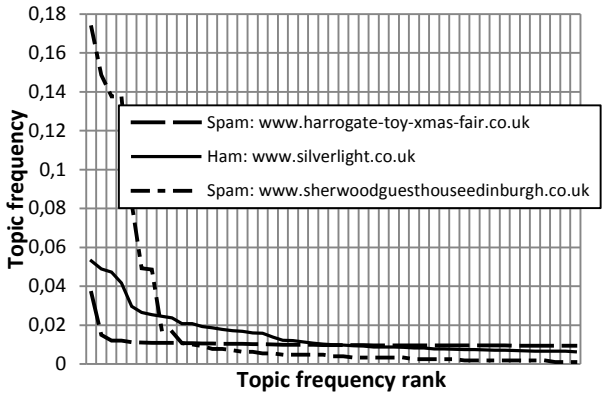

Figure 5. Topical frequency distributions for different types of spam

3.2.3 Sentence Structure Diversity

Most of content spam generation techniques produce new unique texts from a set of natural samples. Spammers may use Markovian text generator, which is trained on a set of natural documents, or they may simply take sentences from different texts, to form a single page content. These techniques often yield locally coherent texts that are hard to detect. To fight these types of spam we developed a set of features to measure the diversity of styles used in text.

We elaborated POS features described in Section 3.1 by adding a wide range of linguistic diversity features to detect style anomalies in texts. For each one of 145 POS ratio features we calculated its variance across sentences of text. A distribution of variances of adjectives ratio is shown in Figure 2, similar distributions work for other parts of speech and different grammatical categories. The graph confirms our hypotheses that content spam tends to mix styles from different texts, resulting in higher variances.

3.3 Topical Analysis

Web spam has a tendency to belong to several popular topics, like insurance, or pornography. We used topical features for two purposes. Firstly we used Latent Dirichlet Allocation (LDA) to measure the weights of different topics in texts and used these weights as classification features. Secondly we analyzed the frequency rank distributions for these weights in order to detect topical structure anomalies.

3.3.1 LDA

We decided to implement a set of topical classification features using Latent Dirichlet Allocation [5]. LDA is a fully generative probabilistic model for texts. LDA assumes that each document is generated by a mixture of topics. The weights of these topics can be used for topical classification. Most importantly LDA weights were used to measure topical diversity of texts. LDA-based topical diversity features are described in Section 3.3.2.

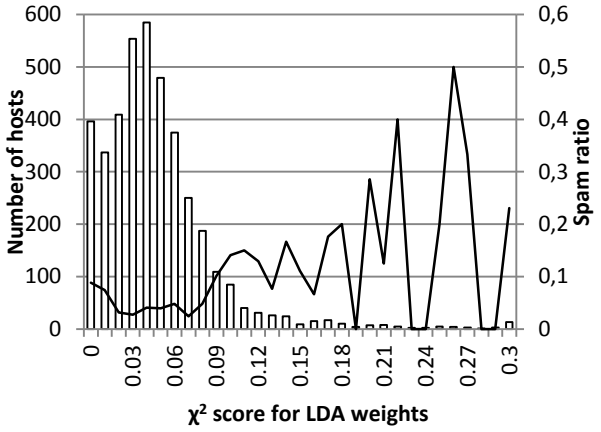

Figure 6. Prevalence of spam relative to chi-squared topical score

LDA has well-established parameter estimation and inference procedures, based on Monte Carlo Markov chains [2]. We used GibbsLDA++ library [18] that implements Gibbs Sampling algorithm for inference and parameter estimation. We trained LDA model on 20K random documents from WEBSPAM-UK2007 dataset, using 100 topics and $\alpha = 0.5, \beta = 0.01$ for hyper-parameters.

We could have used tf.idf for topical classification, but LDA also served as a dimensionality reduction algorithm. As a result we mapped every document in 100-dimension topic space, instead of using high-dimensional term vector space. The weights of different topics served as features in classification.

3.3.2 Topical Diversity

The analysis of LDA topic weights showed that these weights also have a power law distribution. Figure 5 shows the weights distribution for several samples of spam and non-spam hosts. Topical distributions are correlated with term frequency distributions, but have an advantage over them. LDA accounts for correlated terms thus a single LDA topic usually covers a whole set of terms that often co-occur. This ensures that synonyms and similar terms are counted together, and leaves spammers less chances to affect the feature.

For each document we estimated the uniformity of frequency rank distributions of the LDA weights using the formula (*) using topic frequencies instead on word frequencies. The prevalence of spam for different levels of topical uniformity is shown in Figure 4. The probability of spam is greater for hosts with both high and low uniformity. These two zones account for different types of content spam.

Hosts with higher uniformity usually contain texts stuffed with keywords (e.g. www.sherwoodguesthouseedinburgh.co.uk, Figure 3a). The other group of spam hosts has very low topical entropy. Texts from this group usually contain search results or sentences taken from multiple other texts (e.g. www.harrogate-toy-xmas-fair.co.uk, Figure 3b). Topical distributions for these hosts are provided in Figure 5.

We also researched an alternative approach to measuring the topical diversity. Being a probabilistic model LDA only generates the most probable topics weights distribution for the text. In order to detect spam content we calculated the probability of a document having uniform topical distribution (all topics having the same weight). Considering this as a statistical hypothesis the Pearson's chi-squared statistics can be used to check it. Let N be the number of topics, then:

$$\chi^2 = N \sum_{topic} \frac{\left(1/N - weight_{topic}\right)^2}{1/N}.$$

We used this statistics as a classification feature. The prevalence of spam depending on χ^2 score is provided in Figure 6. The higher χ^2 score means that the hosts have lower probability of having uniform topical distribution. The spam probability for hosts with χ^2 score greater than 0.1 is substantially higher than average spam probability.

3.4 Machine Learning

Using LDA as a dimensionality reduction algorithm allowed us to use algorithms designed for dense data, without implementing complex ensembles of classifiers.

We used logistic regression with L2 regularization. We used a fixed regularization parameter value of 0.25. It generates a relatively simple linear classifier with regression coefficients which can be interpreted as contribution of features to the classification task. Some features such as topical uniformity show non-linear behavior that cannot be accounted for using a linear classification formula.

3.5 Complexity estimation

To prove that the proposed algorithm can be used in web-scale spam detection tasks we also estimated the complexity of the proposed algorithm during the classification phase. The algorithm can be loosely split in 3 parts:

- Statistical features calculation;
- Topical diversity estimation based on LDA;
- Machine learning;

The first phase includes POS tagging and compressibility analysis. We used simple POS taggers that analyze single words and do not take previous words in account. The complexity of the POS tagging process in on the order of document's length $O(|d|)$.

The first phase also includes term-level diversity calculation that implies words being sorted by their frequencies. So the complexity of the diversity calculation is on the order of $O(|d|\log(|d|))$.

The second part of the algorithm starts with LDA inference. Gibbs sampling is used for inference and complexity of each iteration is proportional to the length of the document and number of topics used [18]. Instead of running Gibbs sampling until convergence we used fixed number of iterations that suited our purposed well. So the complexity of the Gibbs sampling phase was $O(|d|)$.

Table 1. Feature strength analysis

Feature	F-measure	Feature type
Topical uniformity	91.23%	Diversity
Gzip compression rate	89.70%	Diversity
χ^2 score for LDA weights	87.03%	Diversity
bz2 compression rate	85.04%	Diversity
Term uniformity	81.28%	Diversity
Average number of words repeated in neighbor sentences	79.60%	Diversity
Verbs in past tense ratio	74.49%	Statistical
Average number of expressive punctuation marks per sentence	73.54%	Statistical
Verbs in past tense variance	73.34%	Diversity
Modal verbs variance	72.88%	Diversity
Fraction of sentences with several verbs	71.27%	Statistical
Personal pronouns ratio	71.13%	Statistical
Proper nouns ratio	71.06%	Statistical
Possessive endings variance	70.66%	Diversity
Words with one syllable ratio	70.63%	Statistical
Modal verbs ratio	70.59%	Statistical
Words with two syllables ratio	70.56%	Statistical
Cardinal numbers variance	70.55%	Diversity
Cardinal numbers ratio	70.06%	Statistical
Determiners ratio	69.82%	Statistical

The calculation of topical diversity after the topic weights were estimated depends only on number of topics and its complexity can assumed constant.

Finally in machine learning phase we used constant number of features in a linear classification formula and its complexity is also constant. In whole the complexity of the proposed classification algorithm is $O(|d|\log(|d|))$, where $|d|$ is the length of the classified document.

4 Experiments

The evaluation of the proposed framework consisted of two experiments. First we tested the ability of our approach to detect synthetic automatically generated texts. The second experiment was dedicated to measuring the benefit of the proposed features.

Finally we tested the framework in the Web Spam Challenge [20] settings.

4.1 Synthetic Text Experiment

First we tested the capability of the described features to detect automatically generated low quality texts. We created a set of synthetic texts using a Markovian text generator. The generator was trained on a collection of 20K random documents from the WEBSPAM-UK2007 dataset. Here is a sample of such synthetic text generated from this article:

> *Tf.idf and other term-weighting approaches are often used by web spammers to generate thousands of doorway pages, each optimized for a specific query, to maximize amount of text, and ratio of verbs in past tense compared to all verbs: We used POS tagger to tag every word in the dataset.*

Such texts consist of locally coherent pieces collected from other documents. We used 10K of synthetic documents and random 10K documents from the WEBSPAM-UK2007 dataset as a training set. The test set for the experiment was created in a similar fashion. We used two Markov chains of order 2 (MC2) and 3 (MC3) to measure the effects of this parameter on classification.

In order to measure the effect of the proposed features we made two runs of the experiment. First we used only statistical features and LDA weights as a baseline experiment (SF+LDA). During the second run we used all available features including diversity features (All).

Table 2 contains results of the experiment. High F-measure rate suggests that described features are adequate for detecting such advanced types of content spam. Increase in Markov chain order causes the generator repeat larger pieces of original documents. This reduces detection rate, but increases the amount of non-unique content in such texts. The results also show that the proposed diversity features substantially improve the classifier. In fact they reduce the number of false positives and false negatives in half.

Table 2. Precision, Recall, and F-measure for synthetic text detection experiment

	Precision	**Recall**	**F-measure**
MC2, SF + LDA	96.19%	96.11%	96.15%
MC3, SF + LDA	94.08%	92.29%	93.18%
MC2, All	98.37%	97.93%	**98.14%**
MC3, All	97.72%	97.09%	**97.40%**

Table 3. Results for WEBSPAM-UK2007 experiment

Features	AUC	F1
Geng et. al.	0.85	--
Biro et. al.	0.854	--
SF	0.746	0.284
DF	0.744	0.323
LDA	0.845	0.442
SF+DF	0.777	0.348
SF+LDA	0.867	0.433
DF+LDA	0.864	0.448
All (SF+DF+LDA)	**0.871**	**0.458**

SF – statistical features;
LDA – Latent Dirichlet Allocation topic weights;
DF – diversity features;

4.2 Feature Analysis

The purpose of the second experiment was to estimate power of each of the 334 features. The settings of this experiment were similar to the synthetic text detection experiment. We used 20K documents from WEBSPAM-UK2007 dataset as a non-spam sample and generated 20K documents using a Markov chain text generator with the chain length of 2. These sets were then split evenly in training and testing datasets.

For each feature we trained a separate classifier. Each classifier was trained on a single feature. The classification F-measure of the given classifier can be viewed as a measure of usefulness of the corresponding feature. Table 1 contains the 20 most useful features for synthetic texts classification task.

The results of the experiment show that diversity features are paramount for detecting Markov chain generated texts. The proposed topical diversity features score best on this metric, along with text compressibility. Other diversity features also can be seen among the top-20.

4.3 Webspam-UK2007 Experiment

In this experiment we followed the evaluation protocol of the Web Spam Challenge [20]. Using this evaluation procedure we could compare our results with other studies. The Web Spam Challenge 2008 was held on a WEBSPAM-UK2007 dataset [22]. The training and testing labels are also defined in the dataset. The official quality measure for the challenge was the Area under ROC Curve (AUC ROC). We also calculated optimal F-measure for the classification task.

We compared against best results on this dataset. The winners of the 2008 Web Spam Challenge Geng et. al. [12] used pre-computed features and advanced bagging strategies to reach the AUC of 0.85. Biro et. al. [4] used linked LDA model to combine link and content features yielding the AUC score of 0.854. Dai et. al. [8] used temporal features and achieved classification F-measure of 0.521.

We combined the features into four groups:

- SF – statistical features (Section 3.1);
- DF – various text diversity features (Section 3.2, Section 3.3.2);
- LDA – the Latent Dirichlet Allocation topic weights (Section 3.3.1);

The results of classification using various groups of features and machine learning algorithms are provided in Table 3. Using the logistic regression the best result of 0.871 AUC is achieved when combining all features. Our approach substantially improves over the nearest result of 0.854 AUC. The results show that topical classification features (LDA) are still crucial to web spam detection, but statistical features (SF) and diversity features (DF) improve the results substantially.

5 Conclusion and Future Work

The results of our research show that advanced content features are useful for content spam detection. We analyzed different aspects of natural texts and produced a set of features to cover as many aspects as possible. The resulting spam classifier performed well on both synthetic and real-life tasks.

The proposed approach is based solely on content analysis and doesn't take link data into account. Combining the proposed method with existing link-spam detection techniques is likely to improve results. Another possible extension is to use the diversity measures and rank distributions on link data to detect unnatural link structures.

Web spam is primarily an economic phenomenon and the amount of spam depends on efficiency and costs of different spam generation techniques. We hope that multiple diversity features described in this work can substantially decrease the efficiency of automatically generated content spam. There are many properties of natural texts that are not covered by this article. We plan to continue research on various aspects of natural texts that are hard to reproduce.

References

[1] J. Abernethy, O. Chapelle, and C. Castillo. WITCH: A New Approach to Web Spam Detection. In Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2008.

[2] C. Andrieu, N. de Freitas, A. Doucet, M. Jordan, An introduction to MCMC for machine learning. Machine Learning, 50: 5–43, 2003.

[3] I. Biro, J. Szabo, A. A. Benczur, Latent Dirichlet allocation in web spam filtering, Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web, April 22, 2008, Beijing, China.

[4] I. Biro, D. Siklosi, J. Szabo, A. A. Benczur, Linked latent Dirichlet allocation in web spam filtering, Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web, April 21-21, 2009, Madrid, Spain.

[5] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3(5):993–1022, 2003.
[6] A. Bratko, G. V. Cormack, B. Filipič, T. R. Lynam, and B. Zupan. Spam filtering using statistical data compression models. Journal of Machine Learning Research, 7(Dec):2673–2698, 2006.
[7] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, S. Vigna, A reference collection for web spam, ACM SIGIR Forum, v.40 n.2, p.11-24, December 2006.
[8] N. Dai, B.D. Davison, X. Qi. Looking into the past to better classify web spam. Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web - AIRWeb '09. 2009.
[9] H. Dang. Overview of DUC 2006. Proceedings of the Document Understanding. 2006.
[10] D. Fetterly, M. Manasse, and M. Najork. Detecting phrase-level duplication on the world wide web. In Proceedings of the 28th ACM International Conference on Research and Development in Information Retrieval (SIGIR), Salvador, Brazil, 2005.
[11] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics – Using statistical analysis to locate spam web pages. In Proceedings of the 7th International Workshop on the Web and Databases (WebDB), pages 1–6, Paris, France, 2004.
[12] G. Geng, X. Jin, C.-H. Wang. CASIA at Web Spam Challenge 2008 Track III. In Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2008.
[13] A. Gulin, P. Karpovich. Greedy Function Optimization in Learning to Rank, 2009, Available at: http://romip.ru/russir2009/slides/yandex/lecture.pdf
[14] Z. Gyongyi, H. Garcia-Molina and J. Pedersen. Combating Web Spam with TrustRank. In 30th International Conference on Very Large Data Bases, Aug. 2004.
[15] Z. Gyongyi and H. Garcia-Molina. Web Spam Taxonomy. In 1st International Workshop on Adversarial Information Retrieval on the Web, May 2005.
[16] M. Henzinger, R. Motwani, C. Silverstein. Challenges in Web Search Engines. SIGIR Forum 36(2), 2002.
[17] A. Ntoulas , M. Najork , M. Manasse , D. Fetterly, Detecting spam web pages through content analysis, Proceedings of the 15th international conference on World Wide Web, May 23-26, 2006, Edinburgh, Scotland.
[18] X.-H. Phan, C.-T. Nguyen, Gibbs LDA++: A C/C++ Implementation of Latent Dirichlet Allocation (LDA) using Gibbs Sampling for Parameter Estimation and Inference. http://gibbslda.sourceforge.net/, 2008.
[19] J. Piskorski , M. Sydow , D. Weiss, Exploring linguistic features for web spam detection: a preliminary study, Proceedings of the 4th international workshop on Adversarial information retrieval on the web, April 22, 2008, Beijing, China.
[20] Web Spam Challenge. http://webspam.lip6.fr/wiki/pmwiki.php, 2008.
[21] B. Wu, B. D. Davison. Identifying link farm spam pages. Special interest tracks and posters of the 14th international conference on World Wide Web - WWW '05. 2005.
[22] Yahoo! Research: "Web Spam Collections". http://barcelona.research.yahoo.net/webspam/datasets/ Crawled by the Laboratory of Web Algorithmics, University of Milan, http://law.dsi.unimi.it/. URLs retrieved May 2007.
[23] G. Zipf, Selective Studies and the Principle of Relative Frequency in Language (Cambridge, Mass, 1932).