# WEBSPAM-UK2007 Challenge: Data Analysis School in Moscow

Konstantin Bauman, Alexey Brodskiy, Sergey Kacher, Elmira Kalimulina, Ruslan Kovalev, Mikhail Lebedev, Dmitry Orlov, Pavel Sushin, Pavel Zryumov, Dmitry Leshchiner, Ilya Muchnik

Data Analysis School
1-18 Klimentovsky per
Moscow 115035, Russia

kostuan@mail.ru,
alexey.brodskiy@gmail.com,
serg_kacher@mail.ru,
elmira-yu-k@mail.ru,
ruslan.a.kovalev@gmail.com,
dmblit@gmail.com,
orloff.dmitry@gmail.com,
psushin@gmail.com,
zryumchik@gmail.com,
dmitry_1111@inbox.ru,
muchnik_ilya@yahoo.com

## ABSTRACT

The recognition rule was implemented as a boosted vote of few large margin classifiers built on separate groups of features.

## Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: Hypertext

## Keywords

Web Characterization, Web Spam

## 1. FEATURES

To recognize spam hosts we obtained four types of features for each host:

1) extension and propagation of scores on the host graph,

2) distribution of host page compression rates,

3) word frequencies within host pages,

4) graph topology [2] and page content [3] (features provided by the contest organizers).

The graph features (type 1) were an elaboration of features in [1], extensively reworked due to 10-fold increase of this year's graph size and to the relatively low amount of spam scores available.

The compression features (type 2) were defined as follows: GZIP compression rates for every page of a host were put into 21 bin: [0, 0.5), [0.5, 1), [1, 1.5) ... [9.5, 10), [10, +∞). We made 63 features computing for each bin its page count, average compression rate and standard deviation. We normalized features for the mean=0, std=1 on our training set.

Word frequencies (type 3 features) were computed as follows: frequencies of words in <title>, <meta> (keywords, description), <anchor> and <body> were all counted separately. We computed the percentage of pages containing the word, and the average of log(1+wc)/log(1+pl), where wc – word count, and pl – page length. We also used query log frequencies in word counting for pages. Having frequency distribution in spam and non-spam hosts from the learning set we filtered words with a threshold of 75% discriminating power in Student test.

## 2. LEARNING

The classifier was created by combining weak learners obtained by separate groups of features. There were 13 partial classifiers combined. Combination was done with the TreeNet software [5, 6], in classification mode, with unit weights. While F1 measure of stand-alone classifiers never exceeded 39%, the combined F1 for spam detection reached 67.5% (at 68.3% recall, 66.7% precision).

The partial classifiers were created using SVM with Linear (for type 2 and 4) and Gaussian kernels (for type 4); discriminant functions were then used as weak learners to train TreeNet model. There also was a direct graph-based rule (type 1), and four SVM and three Naïve Bayes classifiers were built on word frequencies.

*AIRWeb'08*, April 22, 2008 Beijing, China.
Copyright 2008 ACM 978-1-60558-159-0…$5.00.

For GZIP: SVMLight [4] with linear kernel was used. Cross-validation figures were R=0.35, P=0.23, F1=0.28. Just 8.63% of non-labeled hosts were classified as spam with these settings.

The work with features provided by organizers was structured as follows. One classifier set has been built using Gaussian kernel SVMLight. The weight –j was set to 1/40. The training set was divided into three equal parts, the first used for SVM training, the other one for kernel gamma parameter tuning, the third one for cross-evaluation. Three different ways of data normalization were used: 1) normalizing features to (mean=0, std=1), 2) normalizing data vectors to |x|=1, 3) the combination of 1) followed by 2). The best achieved F1 (for normalization 3) was 0.39 (R=0.6, P=0.29).

The other classifier used weighted Linear kernel SVMLight with feature selection. For feature selection, the features correlated at level >0.95 were considered connected (and then 276 features yielded 188 connected components). Of each connected component a single feature, most correlated with spam judgments was taken as a representative. That set of 188 features achieved F1=22.4 (R=0.23, P=0.22) with weight of normal class = 0.2.

With the first submission, the combined TreeNet classifier was trained on overall spam judgments provided with judgments by all judges taken together. With the second submission, 34 separate classifiers were built for judgments of each judge that made more than 100 judgments (we took 34 of them), for four judgment classes (borderline, nonspam, spam, unknown). The probabilities of each class were computed for each judge. Then the sum of 34 probabilities for each of first three classes was taken with weight equal to (1-prob("unknown")). Then the final spam probability was calculated as (s + 0.5*b)/(s + n + b), where s, n, b were weighted sums of computed probabilities from all judges for the classes of "spam", "nonspam" and "borderline", respectively.

## 3. ACKNOWLEDGMENTS

The authors are grateful to several people who made this work possible, and especially to Alexander Melkov, Alexei Pyalling, and Sergey Pevtsov, for their invaluable help and contributions.

## 4. REFERENCES

[1] T. Abou-Assaleh, T. Das, 2007. Extention and Propagation of manual Spam scores. (AIRWeb'07, Bannf, Canada.)

[2] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, R. Baeza-Yates: "Using Rank Propagation and Probabilistic Counting for Link-Based Spam Detection". In Proceedings of the Workshop on Web Mining and Web Usage Analysis (WebKDD). Philadelphia, USA, August 2006. ACM Press.

[3] C. Castillo, D. Donato, A. Gionis, V. Murdock, F. Silvestri: "Know your Neighbors: Web Spam Detection using the Web Topology". In Proceedings of ACM SIGIR, pp. 423-430. Amsterdam, Netherlands, 2007. ACM Press.

[4] T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[5] J.H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine." IMS 1999 Reitz Lecture.

[6] T. Hastie, R. Tibshirani and J.H. Friedman, *Elements of Statistical Learning*. Springer, 2001