Web Spam Hunting @ Budapest[*]

Dávid Siklósi András A. Benczúr István Bíró Zsolt Fekete Miklós Kurucz
Attila Pereszlényi Simon Rácz Adrienn Szabó Jácint Szabó
Data Mining and Web search Research Group, Informatics Laboratory
Computer and Automation Research Institute of the Hungarian Academy of Sciences
{sdavid, benczur, ibiro, zsfekete, mkurucz, peresz, sracz, aszabo, jacint}@ilab.sztaki.hu

ABSTRACT

We use a combination, in the expected order of their strength, of the following classificators: SVM over tf.idf, an augmented set of the public statistical spam features, graph stacking and text classification by latent Dirichlet allocation and compression, the latter two only used in our second submission.

1. THE METHOD

We split features into related sets and for each we use the best fitting classifier. These classifiers are then combined by random forest, a method that, in our crossvalidation experiment, outperformed logistic regression suggested by [6]. We used the classifier implementations of the machine learning toolkit Weka [8]. Our expected results obtained by crossvalidation over the training data are shown in [3].

Graph stacking, a methodology used with success for Web spam detection by e.g. [5] is performed under the classifier combination framework as follows. First the base classifiers are built and combined that give prediction $p(u)$ for all the unlabeled nodes u. Next for each node v we construct new features based on the predicted $p(u)$ of its neighbors and the weight of the connection between u and v as described in [7] and classify them by a decision tree. Finally classifier combination is applied to the augmented set of classification results; this procedure is repeated in two iterations as suggested by [5, 7].

In prior results, stacked graphical learning considered edges between the units of the classification, thus the information on the internal linkage as well as the location of an in or outlink within the site structure is lost for the classification process. We used stacked graphical features based on the "Connectivity Sonar" of Amitay et al. [1]. These include the distribution of in and outlinks labeled spam within the site; the average level of spam in and outlinks; the top and leaf level link spamicity.

Content classification quality was improved by adding classifiers based on latent Dirichlet allocation and text compression. In the *multi-corpus LDA* technique [3] we create a bag-of-words document for every Web site and run LDA both on the corpus of sites labeled as spam and as non-spam. In this way collections of spam and non-spam topics are created in the training phase. In the test phase we take the union of these collections, and an unseen site is deemed spam if its total spam topic probability is above a threshold.

Text compression is first used when email spam detection methods applied to Web spam were presented at the Web Spam Challenge 2007 [6]. Similar to [6] we use the method of [4] that compresses spam and nonspam separately; features are defined based on how well the document in question compresses with spam and nonspam, respectively.

Finally we augmented the public challenge features [5] by two features suggested in [2]: the Online Commercial Intention (OCI) value assigned to an URL in a Microsoft adCenter Labs Demonstration as well as measures of how well a page fits to the most competitive queries. Here query competitivity is measured by Google AdWords. In addition we used the following features as well: the number of document formats (`.pdf` etc), the existence and value of `robots.txt` and robots meta; the existence and average of server last modified dates; finally the distance and personalized PageRank from DMOZ sites. We classified by decision trees.

2. REFERENCES

[1] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer. The Connectivity Sonar: Detecting site functionality by structural patterns. In *Proc. 14th ACM Conference on Hypertext and Hypermedia (HT)*, pages 38–47, 2003.

[2] A. A. Benczúr, I. Bíró, K. Csalogány, and T. Sarlós. Web spam detection via commercial intent analysis. In *Proc. 3rd AIRWeb*, 2007.

[3] I. Bíró, J. Szabó, A. A. Benczúr. Latent Dirichlet Allocation in Web Spam Filtering. In *Proc. 4th AIRWeb*, 2008.

[4] A. Bratko, B. Filipič, G. Cormack, T. Lynam, and B. Zupan. Spam Filtering Using Statistical Data Compression Models. *The Journal of Machine Learning Research*, 7:2673–2698, 2006.

[5] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: Web spam detection using the web topology. Proc. SIGIR, 2006.

[6] G. Cormack. Content-based Web Spam Detection. In *Proc. AIRWeb*, 2007.

[7] K. Csalogány, A. Benczúr, D. Siklósi, and L. Lukács. Semi-Supervised Learning: A Comparative Study for Web Spam and Telephone User Churn. In *Proc. Graph Labeling Workshop*, 2007.

[8] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[*]This work was supported by the EU FP7 project LiWA - Living Web Archives and by grants OTKA NK 72845, *ASTOR* NKFP 2/004/05

AIRWeb '08, April 22, 2008 Beijing, China.