Assessing the Quality of Web

Thesis Statements

Dávid Siklósi

Supervisor: András A. Benczúr Ph.D.

Eötvös Loránd University
Faculty of Informatics
Department of Information Systems

Ph.D. School of Computer Science
Erzsébet Csuhaj-Varjú D.Sc.

Ph.D. Program of Information Systems
János Demetrovics D.Sc.

Budapest, 2016.

1 Introduction

In the last decades, Web has grown to be a central part of our lives and substantially changed the way we share information. Users of the Web are not only gathering information from different sources, but also actively editing it via wikis, blogs, forums, social networks and they are commenting, reviewing, tagging and giving opinions on existing content. The immense and continuously growing size of the Web combined with such diversity of information sources makes it nearly impossible to find relevant information without the help of search engines, Web archives and Web directories.

Search engines index billions of Web pages and provide an interface for users to find relevant documents by keyword search. The main task of a search engine is to rank relevant documents in a way that the users have to run through only the first few hits of the result list to find the desired information. In order to retrieve a "good quality" result list, search engines apply sophisticated ranking algorithms based on the keywords the user typed in, the textual content of the Web pages and the Web graph.

In the mid 90's, owners of Web pages started to realize that they gain financial benefits by appearing in the top results of search engines. Thus they started to optimize their pages to influence the ranking algorithms of search engines and attract more and more visitors. While some of these activities help improving the quality of Web pages, a number of shifty webmasters started to stuff their pages with very popular query words and use other techniques to increase their ranking scores for query words that are not even relevant to the content of their Web pages. These pages are called Web spam pages. Since spamming activities deteriorate the quality of search results, it became a top priority goal for search engines to recognize and filter spam.

Besides filtering useless content, search engines also invest huge effort in personalizing their search results. Consider the case when a user types the query: *crane*. It can refer to an animal, a machinery for lifting heavy weights or the American novelist, Stephen Crane. To decide which topic could be the most favorable for the user, search engines categorize the pages of the Web and create profiles of their users' category preferences based on the queries they typed in and the pages they visited earlier.

In my thesis I survey the existing results and introduce some new techniques for automatically identifying spam pages and for classifying the topic of Web pages. I also introduce some new aspects of Web quality, which could help assessing Web pages along more dimensions than spam and topical genre.

2 New Results

Statement 1 Graph Stacking

Several results has appeared that apply rank propagation to extend initial trust [11, 15] or distrust [14, 7] judgments over a small set of seed pages or sites to the entire Web. On the other hand, various link-based algorithms were designed to evaluate node-to-node similarities in networks that can be used to give alternate, similarity based weights to node pairs.

Our semi-supervised, stacked graphical learning algorithm combines the above methodologies with basic Web classification methods as follows:

1. We give prediction for unlabeled data via a machine learning algorithm.

2. For a given unknown node u and edge weight function w, our algorithm selects the k largest weight neighbors of u. We generate a new feature for node u by aggregating the predicted values in this neighborhood.

3. We rerun our machine learning algorithm on the extended feature set and continue our algorithm with step 2

We compared various similarity measures, including simple and multi-step neighborhood, co-citation, cosine and Jaccard similarity of the neighborhood as well as their multi-step variants [9]. We also investigated several aggregation methods to generate the new features.

We ran out experiments on two dataset. For Web spam filtering we used the WEBSPAM-UK-2006 [4] and for churn classification we used data from a small Hungarian landline telephone company.

> **Statement 1.1 Graph stacking methods in combination with graph similarity measures significantly improves classification accuracy both for Web spam filtering and telephone network user churn detection.**

The results are a joint work with András Benczúr, Károly Csalogány and László Lukács and were published in [C1]. My contribution is the implementation of the link feature generation methods, the definition of graph similarity measures, and the identification of the best machine learning methods available as part of the Weka machine learning toolkit. László Lukács defined the aggregation methods for graph stacking and pre-processed the churn dataset. Károly Csalogány calculated the TFIDF feature set and evaluated the results.

Statement 2 Sonar Stacking

A key step in the above described stacked graphical method is turning the predicted labels of the neighborhood into new features of a node by some simple aggregation, such as the majority or average [3] of the neighbor labels. When it comes to host-level web classification, the use of such simple aggregates means that the information about the internal link structure of a host, the exact position of page-level in-links and out-links is lost.

Amitay et al. [1] classify web hosts based on information about their internal link structure. They define the depth of a page within a host as the number of slashes in the corresponding URL and extract features such as:

- the distribution of pages at different levels;
- the distribution of inlinks and outlinks at different levels;
- the fraction of links at the top and leaf levels; and
- comparisons of the distribution of external and internal up, down, and cross links.

We experimented with a refined set of stacked graphical features that combine the predictions of a node's neighbors with information about the internal structure of the node, and the location of links to and from its labeled neighbors. Accordingly, for a host x and some class label c, we extend the "connectivity sonar" features proposed by Amitay et al. by measuring

- the distribution of links to and from neighbor hosts with predicted label c;
- the average level (within the host's internal link graph) of c inlinks and outlinks; and
- the fraction of c links at the top and leaf levels.

We ran our experiments on the WEBSPAM-UK-2007 dataset extended with topical category labels from the DMOZ open directory project.

> **Statement 2.1 By involving the internal structure of Web sites into the aggregation methods of our graph stacking algorithm, we can achieve considerable improvement.**

This enhanced and greatly improved version of graph stacking is a joint work with András Benczúr, Zoltán Gyöngyi and Miklós Kurucz. My main contribution was creating the new sonar stacking features, running all the experiments, comparing different methods for combining the results of separate machine learning algorithms. Zoltán Gyöngyi labeled the WEBSPAM-UK dataset with the DMOZ categories. Miklós Kurucz computed the original sonar features for the baseline methods.

Statement 3 Cross-lingual text classification

While English language training data exists for several Web classification tasks, most notably for Web spam, we face an expensive human labeling procedure if we want to classify Web hosts in a language different from English. Traditional methods in cross-lingual information retrieval use dictionaries, machine translation methods, and more recently multilingual Wikipedia editions. In our methods scalability considerations play central role, hence we cannot rely on heavy NLP machinery that will likely not scale to the Web size. Web classification on the other hand relies on features of content and linkage aggregated over hosts [3], some of which are language independent. However the language independent features are not as strong as bag of words representations [8].

In this work we compared several methods based on language independent features, dictionaries and our new, semi-supervised learning method which exploits hosts that contain a mix of English and national language content. The advantage of this semi-supervised method is that we can use the very strong bag of words features, meanwhile we don't need a dictionary. Our method work as follows:

1. We build our models on the labeled English corpus.
2. We apply our models to the English part of mixed language hosts in the target domain.
3. We use the other part of these hosts, and the predictions we got in the previous step, to create a training data set in the target domain and to build our models in the target language.
4. We apply our new models to the target hosts.

We ran our experiments on a 2009 crawl of the Portuguese Web Archive. For training our English language models, we used the English part of ClueWeb09 dataset. Web spam labels were provided by the Portuguese Web Archive and the Waterloo Spam Rankings [5], respectively. Beside spam, we also labeled hosts in both the `.pt` crawl and ClueWeb09 by top-level DMOZ categories.

Statement 3.1 We demonstrated that the strongest resource for cross-lingual Web classification consists of multilingual Web sites that discuss the same topic in different languages.

Statement 3.2 The normalization of the basic Web spam content features [3] across languages seems to fail for identifying spam and also performs weak for topical classification.

The results are a joint work with András Garzó, Bálint Daróczy, Tamás Kiss and András Benczúr and extend the results published in [C2] by giving improved models and experimenting with ODP categories in addition to Web spam classification. My contribution is running the classification methods on the link and content features, crawling and labeling the Portuguese sites with the DMOZ categories. Tamás Kiss implemented a distributed framework for calculating the link features, András Garzó implemented the hadoop jobs for calculating the content and BM25 features and he did the word by word translation of Portuguese sites. Bálint Daróczy ran the SVM over the BM25 features.

Statement 4 Text classification via bi-clustering

Mining opinion from the Web and assessing its quality and credibility became a well-studied area [6]. Known results typically mine Web data on the micro level, analyzing individual pages of blogs or even sections containing comments and reviews. Instead of relying on the heavy machinery of opinion mining and sentiment analysis [13], our aim is to assess the overall quality of hosts, at Web scale.

The ECML/PKDD Discovery Challenge 2010 on Web Quality (DC2010) introduced the tasks for assessing neutrality, bias and trustworthiness of Web hosts. We consider these attributes as key aspects of Web quality. However, participants of DC2010 [10, 2, 12] found the new quality tasks particularly challenging with AUC values, in all cases, below 0.6, typically even near the 0.5 value of a completely random prediction.

As the bag of words representation turned out to describe Web hosts best for most classification tasks of the Discovery Challenge [8], we realized that new text classification methods are needed particularly suited to the quality related tasks in question and we came up with the following solution:

1. Instead of using bag of words representation, we compile a bag of concepts representation from words via bi-clustering. This low dimensional representation allows computationally costly classifiers, in particular SVM, to be applied.

2. We use simple feature selection and weighting based on frequencies in the training set. Unlike for all other categories of spam and genre, this method is particularly suited to the highly imbalanced classes of non-neutrality, bias and distrust.

3. Given the compact representation of hosts by cluster distances, we may apply computationally expensive methods for classification. We use SVM with several kernels and we use early and late kernel fusion to combine them.

We compared our results over the DC2010 data set, both with the best results of the participants [10, 2, 12] and with our earlier results focusing primarily on spam classification [8]. We also evaluated

our classification techniques on the C3 data set of the WebQuality 2015 Data Challenge.

Statement 4.1 Our Bag of concepts based method combined with the known baseline methods significantly improves the results in Web quality classification.

The experiments on the DC2010 dataset are a joint work with Bálint Daróczy, András Benczúr and were published in [C3]. For the experiments on the C3 dataset Róbert Pálovics was also involved and our results on were published in [C4]. My main contribution is the design and implementation of the bi-clustering algorithm that can be used for text classification. The code is available on github[1]. In addition, I crawled and parsed the pages of the C3 data set, constructed the textual features and evaluated all the results. Róbert Pálovics ran the matrix factorization and gradient boosted tree baseline methods on the C3 data set. Bálint Daróczy experimented with the multiple kernel methods.

References

[1] Einat Amitay, David Carmel, Adam Darlow, Ronny Lempel, and Aya Soffer. The Connectivity Sonar: Detecting site functionality by structural patterns. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia (HT)*, pages 38–47, Nottingham, United Kingdom, 2003.

[2] Ludovic Denoyer Artem Sokolov, Tanguy Urvoy and Olivier Ricard. Madspam consortium at the ecml/pkdd discovery challenge 2010. In *Proceedings of the ECML/PKDD 2010 Discovery Challenge*, 2010.

[3] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: web spam detection using the web topology. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 423–430, 2007.

[4] Carlos Castillo, Debora Donato, Luca Becchetti, Paolo Boldi, Stefano Leonardi, Massimo Santini, and Sebastiano Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, December 2006.

[5] G.V. Cormack, M.D. Smucker, and C.L.A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval*, 14(5):441–465, 2011.

[1] https://github.com/siklosid/co-cluster.git

[6] K. Dave, S. Lawrence, and D.M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.

[7] Isabel Drost and Tobias Scheffer. Thwarting the nigritude ultramarine: Learning to identify link spam. In *Proceedings of the 16th European Conference on Machine Learning (ECML)*, volume 3720 of *Lecture Notes in Artificial Intelligence*, pages 233–243, Porto, Portugal, 2005.

[8] Miklós Erdélyi, András Garzó, and András A. Benczúr. Web spam classification: a few features worth more. In *Joint WICOW/AIRWeb Workshop on Web Quality (WebQuality 2011) In conjunction with the 20th International World Wide Web Conference in Hyderabad, India.* ACM Press, 2011.

[9] Dániel Fogaras and Balázs Rácz. Scaling link-based similarity search. In *Proceedings of the 14th World Wide Web Conference (WWW)*, pages 641–650, Chiba, Japan, 2005.

[10] Xin-Chang Zhang Guang-Gang Geng, Xiao-Bo Jin and Dexian Zhang. Evaluating web content quality via multi-scale features. In *Proceedings of the ECML/PKDD 2010 Discovery Challenge*, 2010.

[11] Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, pages 576–587, Toronto, Canada, 2004.

[12] Vladimir Nikulin. Web-mining with wilcoxon-based feature selection, ensembling and multiple binary classifiers. In *Proceedings of the ECML/PKDD 2010 Discovery Challenge*, 2010.

[13] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[14] PR10.info. BadRank as the opposite of PageRank, 2004. `http://en.pr10.info/pagerank0-badrank/` (visited June 27th, 2005).

[15] Baoning Wu, Vinay Goel, and Brian D. Davison. Topical TrustRank: Using topicality to combat web spam. In *Proceedings of the 15th International World Wide Web Conference (WWW)*, Edinburgh, Scotland, 2006.

Publications

[C1] Károly Csalogány, A.A. Benczúr, D. Siklósi, and L. Lukács. Semi-Supervised Learning: A Comparative Study for Web Spam and Telephone User Churn. In *Graph Labeling Workshop in conjunction with ECML/PKDD 2007*, 2007.

[C2] András Garzó, Bálint Daróczy, Tamás Kiss, Dávid Siklósi, and András A Benczúr. Cross-lingual web spam classification. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1149–1156. International World Wide Web Conferences Steering Committee, 2013.

[C3] D. Siklósi, B. Daróczy, and A.A. Benczúr. Content-based trust and bias classification via biclustering. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 41–47. ACM, 2012.

[C4] Bálint Daróczy, Dávid Siklósi, Róbert Pálovics, András A. Benczúr. Text classification kernels for quality prediction over the C3 data set. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 1441?1446. International World Wide Web Conferences Steering Committee, 2015.

Other Related Publications

[P1] András Benczúr, István Bíró, Mátyás Brendel, Károly Csalogány, Bálint Daróczy, and Dávid Siklósi. Multimodal retrieval by text–segment biclustering. In *Advances in Multilingual and Multimodal Information Retrieval*, pages 518–521. Springer, 2008.

[P2] András A. Benczúr, Miklós Erdélyi, Julien Masanés, and Dávid Siklósi. Web spam challenge proposal for filtering in archives. In *AIRWeb '09: Proceedings of the 5th international workshop on Adversarial information retrieval on the web*. ACM Press, 2009.

[P3] András A. Benczúr, Dávid Siklósi, Jácint Szabó, István Bíró, Zsolt Fekete, Miklós Kurucz, Attila Pereszlényi, Simon Rácz, and Adrienn Szabó. Web spam: a survey with vision for the archivist. In *Proc. International Web Archiving Workshop*, 2008.

[P4] István Bíró, Jácint Szabó, and András A. Benczúr. Latent Dirichlet Allocation in Web Spam Filtering. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[P5] Bálint Daróczy, Zsolt Fekete, Mátyás Brendel, Simon Rácz, András Benczúr, Dávid Siklósi, and Attila Pereszlényi. Cross-modal image retrieval with parameter tuning. In Carol Peters, Danilo Giampiccol, Nicola Ferro, Vivien Petras, Julio Gonzalo, Anselmo Peñas, Thomas Deselaers, Thomas Mandl, Gareth Jones, and Nikko Kurimo, editors, *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, September 2008 (printed in 2009).

[P6] Bálint Daróczy, István Petrás, András A Benczúr, Zsolt Fekete, Dávid Nemeskey, Dávid Siklósi, and Zsuzsa Weiner. Interest point and segmentation-based photo annotation. In *Multilingual Information Access Evaluation II. Multimedia Experiments*, pages 340–347. Springer, 2010.

[P7] Bálint Daróczy, Dávid Siklósi, and András A Benczúr. Dms-sztaki@ imageclef 2012 photo annotation. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[P8] Miklós Erdélyi, András A Benczúr, Bálint Daróczy, András Garzó, Tamás Kiss, and Dávid Siklósi. The classification power of web features. *Internet Mathematics*, 10(3-4):421–457, 2014.

[P9] Miklós Erdélyi, András A. Benczúr, Julien Masanés, and Dávid Siklósi. Web spam filtering in internet archives. In *AIRWeb '09: Proceedings of the 5th international workshop on Adversarial information retrieval on the web*. ACM Press, 2009.

[P10] M. Kurucz, L. Lukács, D. Silklói, A.A. Benczúr, K. Csalogány, and A. Lukács. Telephone call network data mining: A survey with experiments. *Handbook of Large-Scale Random Networks*, pages 489–530, 2008.

[P11] Miklós Kurucz, Dávid Siklósi, István Bíró, Péter Csizsek, Zsolt Fekete, Róbert Iwatt, Tamás Kiss, and Adrienn Szabó. Kdd cup 2009 @ budapest: feature partitioning and boosting. *Journal of Machine Learning Research special issue on KDD Cup and Workshop in conjunction with KDD 2009*, 2009.

[P12] Miklós Kurucz, Dávid Siklósi, Károly Csalogány, László Lukács, András Benczúr, and András Lukács. Kapcsolatok és távolságok: a hazai vezetékes hívás-szokások elemzése. *Magyar Tudomány*, 170(6):697–706, 2009.