Developing Intelligent Search Engines

Isabel Drost

Abstract

Developers of search engines today do not only face technical problems such as designing an efficient crawler or distributing search requests among servers. Search has become a problem of identifying reliable information in an adversarial environment. Since the web is used for purposes as diverse as trade, communication, and advertisement search engines need to be able to distinguish different types of web pages. In this paper we describe some common properties of the WWW and social networks. We show one possibility of exploiting these properties for classifying web pages.

1 Introduction

Since more and more data is made available online, users have to search an ever growing amount of web pages to find the information they seek. In the last decade search engines have become an important tool to find valuable web sites for a given query. First search engines did rely on simply computing the similarity of query and page content to find the most relevant sites. Today, most engines incorporate some external relevance meassure, like the page rank [23], to determine the correct ranking of web pages. Intuitively, each web page gets some initial "page rank", collects additional weight via its inlinks and evenly distributes the gathered rank to all pages it links to. Thus, pages that have either many inlinks from unimportant pages or at least some links from sites already considered important are ranked high.

Nowadays the WWW is not only used to publish information or research results as was done in its very beginning. Many web pages we encounter are created to communicate, trade, organise events or to promote products. The question arises whether it is possible to identify certain types of web pages automatically and augment the search results with this information. In this paper we investigate certain properties of the web graph that can also be found in social networks. These properties distinguish natural occuring graphs from synthetic ones and can for instance be used to identify link spam [14].

The paper gives a short overview of how the link graph can be used to distinguish certain types of web pages with machine learning techniques. Basic notation conventions are introduced in chapter 2. An overview of different link graph properties is given in 3, chapter 4 deals with the classification of different web page types. Open problems are presented in chapter 5.

2 Notation

The world wide web can be represented as a graph $G = V, E$. Each page corresponds to one node (also referred to as vertice) v_i in the graph. Each link e_{ij} from page i to j is represented as an edge. The outdegree of v_i corresponds to the number of links originating from this node (outlinks), its indegree to the number of links pointing to this page (inlinks).

We refer to pages linking to v_i, as well as those linked by v_i by the term link neighborhood.

3 WWW as Social Network

Social networks such as graphs representing relationships between humans or biological constraints can be shown to differ from synthetic networks in many properties [18, 22, 24]. In the following we shed light on a selection of differences that can be exploited to distinguish different types of web pages.

3.1 Power Law Distribution

The distribution of the nodes' indegree in social networks is power law governed [12, 2]. Intuitively speaking this means, that a large proportion of nodes have few links pointing to them whereas there are only a few pages that attract large amounts of links.

The indegree distribution of web pages should also be exhibit a power law. Yet in [12] empirical studies have shown that the actual distribution has several outliers. Examining this problem more deeply, the authors found the outliers being link spam in most of the cases. They concluded that this observation might help in designing a link spam classifier.

3.2 Clustering Coefficient

Given a node x_i in an undirected graph, the clustering coefficient of this node gives the proportion of existing links among its neighbors vs. all links that theoretically could exist. In [18] Newman observed that this coefficient is higher in naturally occuring networks than in synthetic networks: Two nodes both linked to a third one are also linked to each other with high probability in naturally occuring networks.

For the WWW no clear decission could be made on whether its average clustering coefficient is similar to the one in natural networks. In our work however [10] we observed that the local clustering coefficient can be exploited successfully to distinguish spam from ham web pages. The local clustering coefficient simply gives the probability of a link between two randomly drawn link neighbors of one web page.

3.3 Small World Graphs

The most commonly known property of small world graphs is, that the shortest path connecting two nodes is considerably smaller than in random graphs. The concept became widely known after Milgrams publication [20] that suggested that US citizens are on average connected to each other by a path of 6 intermediate nodes.

The WWW also should reveal such properties. But what we observe in reality is a deviation from these statistics: Large networks of link spam deteriorate the small world properties of the WWW [7, 6].

4 Classifying Web Pages

In this section we treat the problem of classifying web pages. We incorporate the local link structure of the example nodes into their feature representation. To validate the expressivity of this representation we apply this strategy for classifying link spam.

4.1 Representing Examples

Each web page is represented by three feature sets. The first one corresponds to intrinsic properties such as the length of the example page to classify. The second set covers averaged and summed features of the link neighborhood, such as the average length of pages linking to the example to classify. The last set of features covers the relations among neighboring pages and the example such as the average number of pages among the inlink pages with same length as the example vertice. A detailed description of the features used is given in table 1.

Many of these features are reimplementations of, or have been inspired by, features suggested by [9, 10] and [11].

4.2 Classifying Web Spam

In our experiments, we study learning curves for the tfidf representation, the attributes of Table 1, and the joint features. Figure 1 shows that for all spam, combined features are superior to the tfidf representation.

In [10], a ranking of features according to their importance for classification is given. The most important features all cover attributes of either the neighboring pages themselves as well as context similarity features. Also the clustering coefficient of the pages to classify ranges rather high in the feature ranking. This publication also examines the behavior of the classifier in an adversarial environment: Spammers adopt the structure of their web pages as soon as a new spam classifier is employed. The experiments show that our link spam classifier needs to be retrained quickly, in case spammers start to adopt their web pages.

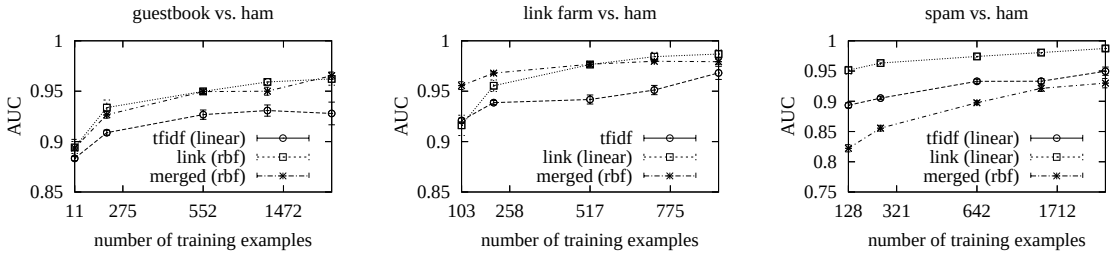

Figure 1: Comparison of feature representations.

5 Open Problems

5.1 Learning for Search Engines

In order to improve the user interface of search engines, there has been large effort in developing algorithms that are able to dynamically cluster search results for a given query [21, 25]. Presenting clusterd search results seems especially appealing when dealing with ambigous queries such as the term "cluster". Today there already exist many search engines that actually use clustering algorithms to present their search results. Nevertheless, the clusters generated as well as their descriptions are far from perfect.

In [15] Henzinger identifies many open problems in web search itself. According to her, one of the most important and challenging research area is the automatic identification of link spam as well as the identification of duplicate and near duplicate web pages.

Recently there have been many publications on the topic of web spam. Both the problem of creating optimal link farms [4, 1] as well as the problem of actually identifying link spam [5, 26, 10] were adressed. Yet the autocorrelation of labels of neighboring nodes has so far not been taken into consideration when classifying link spam. Work that has been done in the field of collective classification [13, 16] might be a start for the development of link spam algorithms.

5.2 Game Theory

When developing classifiers for the problem of web spam identification the problem of adversarial classification [8] has to be taken into account. As soon as the classifier is implemented into a search application spammers will probe its algorithm and develop new spamming techniques probably unknown to the classifier. So the algorithm's stability against adversarial obfuscation of web pages has to be examined when proposing a new spam classifier.

The problem of web spam could equally well be modelled as a game between spam filter and spammer: The web spammer tries to trick the filter and place the spammed web page as high as possible, the filter tries to identify the spammer vs. regular web content. The question that arises then might be, whether it is possible to prove some kind of equilibrium for this kind of game.

5.3 Other Types of Spam

Click spamming is a particularly vicious form of web spam. Companies allocate a fixed budged to sponsored links programs. The sponsored link is shown on web pages related to the link target. For each click on the link the company has to pay a small amount of money to the enterprise hosting it. Rivaling companies as well as hosting companies now employ "click bots" that automatically click on their competitor's sponsored link and cause massive financial damage.

This practice undermines the benefit of the sponsored link program, and enterprises offering sponsered link programs such as Google therefore have to identify whether a reference to a sponsored link has been made by a human, or by a "rogue bot". This classification task is extremely challenging because the HTTP protocol provides hardly any information about the client.

Table 1: Attributes of web page x_0.

Textual content of the page x_0; tfidf vector.
Features are computed for $X = \{x_0\}$, for predecessors $X = pred(x_0)$, successors ($X = succ(x_0)$). The attributes are aggregated (sum and average) over all $x_i \in X$.
Number of tokens keyword meta-tag.
Number of tokens in title.
Number of tokens in description.
Is the page a redirection?.
Number of inlinks of x.
Number of outlinks of x.
Number of characters in URL of x.
Number of characters in domain of x.
Number of subdomains in URL of x.
Page length of x.
Domain ending ".edu" or ".org"?
Domain ending ".com" or ".biz"?.
URL contains tilde?.
The context similarity features are calculated for $X = pred(x_0)$ and $X = succ(x_0)$; sum and ratio are used for aggregation.
Clustering coefficient of X.
Elements of X with common IP.
Elements of X of common length.
Pages that are referred to in x_0 and also in elements of X.
Pages referred to from two elements of X.
Pages in X with comon MD5 hash.
Elements of X with IP of x_0.
Elements of X with length of x_0.
Pages in X with MD5 of x_0.

5.4 Novelty Detection

The WWW provides a large amount of information that is regularly updated. In many cases, news - such as reports from the 2004 tsunami in south east asia - spread first via private web pages before common news papers are able to write about the event. The detection of new trends, stories and preferences among the huge amount of web pages is an especially challenging task that might influence not only the development of the web itself but also the articles published in news papers.

There are already several publications on the topic of tracking specific topics [3] as well as the identification of new trends [19] in a linked environment. Yet some more work might be necessary to make these ideas work on the large linked graph of web pages.

5.5 Personalized Ranking

At the moment search engines in general employ exactly one ranking function for each query. Unfortunately for users this means that the ranking of search results is a mere compromise of the needs of all the users of the engine. A personalized ranking here might help to find exactly what the user needs: An astronomer searching for the term cluster might be unlikely to seek information on the topic of search result clustering. He probably will search information about clusters of stars. On approach to solve this problem is to provide a ranking function that adopts to the searcher.

Recently a rather exhaustive user [17] study showed, that the users' clicks on search results could be used as implicit feedback on the quality of the ranking. On the basis of these results one could imagine to build a personalized ranking function for each individual user that exactly takes into account which search results the user preferably clicked on in the past.

References

[1] S. Adali, T. Liu, and M. Magdon-Ismail. Optimal link bombs are uncoordinated. In *Proc. of the Workshop on Adversarial IR on the Web*, 2005.

[2] Lada A. Adamic. The small world web. In S. Abiteboul and A.-M. Vercoustre, editors, *Proc. 3rd European Conf. Research and Advanced Technology for Digital Libraries, ECDL*, number 1696, pages 443–452. Springer-Verlag, 1999.

[3] James Allan. Introduction to topic detection and tracking. pages 1–16, 2002.

[4] R. Baeza-Yates, C. Castillo, and V. López. Pagerank increase under different collusion topologies. In *Proc. of the Workshop on Adversarial IR on the Web*, 2005.

[5] A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher. Spamrank – fully automatic link spam detection. In *Proc. of the Workshop on Adversarial IR on the Web*, 2005.

[6] K. Bharat, B. Chang, M. Henzinger, and M. Ruhl. Who links to whom: Mining linkage between web sites. In *Proc. of the IEEE International Conference on Data Mining*, 2001.

[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proc. of the International WWW Conference*, 2000.

[8] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma. Adversarial classification. In *Proc. of the ACM International Conference on Knowledge Discovery and Data Mining*, 2004.

[9] B. Davison. Recognizing nepotistic links on the web, 2000. In Proceedings of the AAAI-2000 Workshop on Artificial Intelligence for Web Search.

[10] Isabel Drost and Tobias Scheffer. Thwarting the nigritude ultramarine: learning to identify link spam. In *Proc. of the ECML*.

[11] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In *Proc. of the International Workshop on the Web and Databases*, 2004.

[12] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *Proc. of the International WWW Conference*, 2003.

[13] Lise Getoor. Link-based classification. Technical report, University of Maryland, 2004.

[14] Zoltn Gyngyi and Hector Garcia. Web spam taxonomy. In *Proc. of the Workshop on Adversarial IR on the Web*, 2005.

[15] M. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. In *Proc. of the International Joint Conference on Artificial Intelligence*, 2003.

[16] David Jensen, Jennifer Neville, and Brian Gallagher. Why collective inference improves relational classification. In *KDD '04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 593–598. ACM Press, 2004.

[17] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005.

[18] Juyong Park M. E. J. Newman. Why social networks are different from other types of networks. Technical report, arXiV cond-mat/0305612, 2003.

[19] Naohiro Matsumura, Yukio Ohsawa, and Mitsuru Ishizuka. Discovery of emerging topics between communities on www. In *WI '01: Proceedings of the First Asia-Pacific Conference on Web Intelligence: Research and Development*, pages 473–482, London, UK, 2001. Springer-Verlag.

[20] Stanley Milgram. The small world problem. *Psychology Today*, 61, 1967.

[21] Dharmendra S. Modha and W. Scott Spangler. Clustering hypertext with applications to web searching. In *ACM Conference on Hypertext*, pages 143–152, 2000.

[22] M. E. J. Newman. Assortative mixing in networks. Technical report, arXiV cond-mat/0205405, 2002.

[23] L. Page and S. Brin. The anatomy of a large-scale hypertextual web search engine. In *Proc. of the Seventh International World-Wide Web Conference*, 1998.

[24] L. Tsimring and R. Huerta. Modeling of contact tracing in social networks. *Physika A*, 325:33–39, 2003.

[25] Yitong Wang and Masaru Kitsuregawa. Link based clustering of Web search results. *Lecture Notes in Computer Science*, 2118, 2001.

[26] Baoning Wu and Brian D. Davison. Identifying link farm spam pages. In *Proc. of the 14th International WWW Conference*, 2005.