One Way to Detecting of Link Spam

Ruslan V. Sharapov, Ekaterina V. Sharapova

Murom Institute of Vladimir State University, Russia, Murom
mivlgu@mail.ru

Abstract: In article is considered the method of link spam detection. The basic place in work is detection of “paid” links as a kind of link spam. Here are analysed the significant characteristics of “paid” links. Based on this information algorithm of link spam detection is described. Finally, in article results of algorithm working is given.

1 Introduction

The quantity of information in Internet becomes more and more every year. The number of sites constantly grows, the number of their pages are increases. For example, Yandex search engine searches more than 8 billion pages, Google – more than 1 trillion pages. The Internet is means of information reception and dialogue, and also it is means for business. The first position of site‘s in list of search results is an actual problem for the site owners. There are many sites of same subjects. That is why everyone wishes to be at top of search results list in queries. For this tasks site owners use spam of search engines (web spam). There are many methods, used for web spam. We will consider one of them – a link spam.

The increasing links number at sites became one of basic methods deceiving of search engines. Scales of this manipulation constantly grow. Several years ago the basic way of link spam was a manual links exchange. Today it is used the various ways of links automatic placing. There are some variants of such placing:

1. Using of specialised programs for automatic addition of links in catalogs, guest books, forums etc.

2. Purchases of links at links (advertising) brokers.

Search engines fight with first variant of spam placing by revealing of resources with possibility simple, not moderated addition of links. The weight of links from such resources strongly decreases. Placing of links with using of links brokers makes big problems for search engines.

Today just in Russian segment of Internet worked about ten the large advertising brokers selling text links. Only one of them, Sape.ru, has possibility to place the links on more than 200 million pages. That links named “advertising”. But it is not true. Links take places often in the most imperceptible places of page and their task is improvement of position in search engines. Cost of such “advertising” also happens often nominal, sometimes only 0.01$ for a one month of placing.

The links placed by means of advertising brokers, we will named “paid” links, underlining that links have an unnatural origin (i.e. have no relation to page content). They take places by owner of page for money, instead of with “respect” for a site to which refer. That is why paid links are spam links.

In what the basic danger of the large-scale link spam, observed at last time? Danger consists in that, what that links are actively used by modern search engines for ranging of search results. With links it is connected the concepts of the Quote Index in Yandex and definition PageRank in Google. The mass increase unnatural origin links (link spam) can strongly “spoil” an overall performance of search systems. The situation becomes complicated that "paid" links can take places on any sites including very dear and popular resources. Thus, there is impossible a simple division of pages on “good” and pages for link spam.

2 Related works

Many works are devoted to link spam detection.

In work [FMN04] the statistical analysis for detecting of automatically generated spam pages is offered. To spam can testify: deviation from normal distribution of various pages properties, such as names and IP-addresses, in and out-links, page content and norm of change.

Many works are devoted the analysis of link information – interrelations of pages united by links and texts of links. Some of developers are offered by algorithms, constructed on PageRank.

In paper [KR06] algorithm Anti-Trust Rank is considered. The algorithm is based on manual selection of pages with and without spam. The further analysis of web-graph constructed on basis of link structures, allows to detect pages using spam. The algorithm shows high accuracy of spam detection, including pages with high PageRank value.

In paper [BCSU05] algorithm SpamRank is offered. The algorithm is based on concept personalised PageRank. SpamRank detect pages with not deserved high PageRank value without use of white or black lists or other means of person intervention.

In work [GGP04] algorithm TrustRank is described. Principle TrustRank based on thesis that “good” pages usually refer to “good” pages and seldom use links for spam. At first set of “good” pages undertakes and it appoints high weight. Further the approach similar

PageRank is used: weight is divided into out-links to other pages. At last, after convergence, pages with high weight are accepted to good pages. Authors consider that using of algorithm TrustRank yields better results, than PageRank.

In work [WD05] the web-graph for detection link spam (link farms) is offered to analyse. The algorithm is based on analysis of in and out-links. If crossing of in and out-links more than certain threshold that to pages is found penalty is appointed. This operation is carried out for all pages.

In work [EMT04] algorithm HostRank (PageRank, calculated on host-graphs) is offered. The algorithm allows reducing number of doubtful sites in search results reducing weight, received by sites from a link spam.

In work [CDG+07] decisions tree C4.5 is applied. Two groups of properties are used. Link properties include: 16 properties of affinity degree (in and out-links, number of cross–links etc.), 11 properties based on PageRank (various measures connected with PageRank of page and PageRank from pages referring to it), Truncated PageRank etc. 24 properties depending on page content. The algorithm offered by authors allows detecting 88,4 % of a spam.

Work [Da00] is devoted detection of link spam. The problem is reduced to splitting of pages into two classes – “spam” and “ham” (not spam). The support vector machines (package SVM-light with standard parameters) is used. For each page 89 properties and TF-IDF vector are allocated. Authors use following properties for classification: number of words in keyword, description meta-tags and title, redirect on page, number of in and out-links, number of symbols in URL and the domain, number of subdomains in URL, size of page, domains zones .edu .org .biz .com, identical IP-addresses, identical size of pages etc. For detection of identical pages authors suggest to use MD5 hash-codes. Authors analyse various variants of kernels SVM and reveal the most important properties for classification.

The analysis shows, that existing algorithms are based on the analysis of links network structure, detecting of spam-pages and sites etc. But they practically are not intended for detection “good” and “spam” links on each separate page.

3 Detecting of Link spam

Let's consider signs for paid links detection:

a) The links noted as advertising

For this purpose it is necessary to see link environment (text near to link). Signs of paid link are words: “Advertising”, “Sponsors”, “Our Partners”, etc.

b) The big block of links.

The big links number on a small page area (block of links) can testify to their unnatural origin.

c) Links to advertising broker / advertising agency.

Often near to advertising blocks it is possible to see links to advertising brokers (links brokers).

d) On a site there is information on how it is possible to buy links.

If on site or about the block of links contains such information, then links are paid.

e) Thematic similarities of link.

If text of link or subject of linked site differs from page subjects on which the link is placed then the link is a spam.

f) Thematic similarities of near links.

For this purpose it is necessary to analyse subjects of group links placed on page. If links are not thematic and have disorder of subjects, they are advertising.

g) The location of links.

For this purpose it is necessary to analyse an arrangement of links on page. Than further the link from the basic maintenance of page, it is especially probable, that they are advertising. For example, often such links take places in the bottom of page or in the right column when the body text placed on the middle of a page.

h) Code of links.

Many automated systems of links placing (link exchange and brokers) establish code automatically on template. Presence of identical links block on code can specify in their spam origin.

i) Dynamics/time of links life.

Frequent change of links on pages without change of other maintenance can testify to their unnatural origin. Links can disappeared during some time from pages (in case of malfunctions of systems on automatic placing of links), or their part can be replaced with new links.

j) The message on paid links.

Competitors, the former buyers of links, the former employees can inform on paid links.

k) Viewing of page by person.

Viewing of pages by a moderator and detect link spam manually.

4 Algorithm of link spam detection

Now we will consider the algorithm, capable to detect spam links. It consists of several stages.

Stage 1: Creation of a preset of spam links *S*.

The set is formed of following links:

- chosen manually;
- defined by algorithm early, as spam;
- defined by analysis of advertising brokers code.

The greatest interest represents last way. Some advertising brokers have distinctive features in placing of codes that could help to identify of them [SS08]. At work of algorithm first two ways can be passed. This is necessary for simplification of search known earlier spam-links.

It is necessary to notice, what not all links defined by algorithm as a spam, should be brought in set *S*. Just links with expressive characteristics of spam can be including in set *S*. It is necessary to exclude casual hit of links in the spam category.

After work of this stage the set *S* can be empty.

Stage 2: Detection of spam links on basis of page content.

The basic idea consists in the analysis of page content and detecting of spam signs. For each spam sign at link it is established a penalty q_i to it.

Step 1. The page is scanned on presence of links S_b, put into list *S* generated at the Stage 1. If such link presents on page then the area around them is scanned. If other links present on the page its sets penalty q_1 which value decreases in process of removal from link S_b.

Step 2. The page is scanned on presence of advertising block signs. As a signs can serve words “Advertising”, “Sponsors”, “Our Partners” etc. If this signs present on the page then to links around it is established penalty q_2.

Step 3. The page is scanned on presence of links to advertising broker. If such block present on page then to links around it is established penalty q_3.

Step 4. The page is scanned on information about sale of links (and about how they can be bought). If such signs present on the page then to links in their vicinities is established penalty q_4.

Step 5. Page is scanned on location on big block of links. If number of links in block more than a certain threshold, its sets the penalty q_5.

Step 6. Links are scanned on signs of advertising broker code. If signs present to links are established penalty q_6.

Step 7. Subject of link and general subject of page are checked. If subjects are different, then to link is set penalty q_7. For checking a subjects it is often enough to scan the page text on coincidence of words to the text of link.

Step 8. Subject of link and subjects of around links is checked. If subjects are different, the link is set penalty q_8.

Step 9. The link location is considered. If the link is in the end of page, it is established a penalty q_9.

This stage is core for this algorithm.

Stage 3. Analysis of site structure for purpose of spam detection.

This stage is the most difficult. Its purpose is revealing features of site structure and finding place on pages with “paid” links.

For this purpose from site pages all changing content (except links) remove. Further association of pages with an identical template in clusters is made. The following stage: for each cluster repeating links remove and areas where links vary on everyone cluster pages are identified. For the links entering into such areas is established penalty q_r.

Stage 4. For each link all added penalty are summarised.

If total sum of penalty is more than some threshold then link is spam. In this case the link is put into list *S*.

All steps of stages 2–4 executed fully automatically. The values of penalty $q_1 - q_9, q_r$ are set before algorithm work [SS08].

5 Results

For formation of spam-links initial set *S* has been scanned 20 sites placing link spam (information about link spam has been given us by site owners). Numbers of pages on each site are from 100 to 5000. After removal of duplicates the set from 15000 spam links has been received.

Algorithm work was estimated by a method [BCSU05].

$$\text{Precision} = \frac{\text{Number of spam links, classified as spam}}{\text{Number of links, classified as spam}}$$

$$\text{Recall} = \frac{\text{Number of spam links, classified as spam}}{\text{Number of spam links}}$$

$$\text{FalseSpam} = \frac{\text{Number of not spam links, classified as spam}}{\text{Number of not spam links}}$$

$$\text{FalseNSpam} = \frac{\text{Number of spam links, classified as not spam}}{\text{Number of spam links}}$$

For testing 100 pages with number of out-links from 1 to 30 on everyone have been manually selected. The total of links was 783. For each page spam links have been manually selected (total 519 links). As result of algorithm work 488 links have been noted as spam, their which 461 really were spam links. Results of algorithm work are resulted in table 1.

Precision	0.94
Recall	0.89
FalseSpam	0.102
FalseNSpam	0.112

Table 1: Results of algorithm testing

The algorithm operating time is some seconds on each page.

6 Conclusions

Offered algorithm shows good results in detection of spam links. The Precision value is 0.94, Recall value is 0.89. It is good result in comparison with other works [BCSU05, Da00, GGP04 etc.]. Developed algorithm is capable to detect “good” and “spam” links on each separate page.

Many errors in detecting of spam links it is caused with approached analysis of thematic similarities. Some links have been detecting as spam (group of links) that part of links in group has been recognised as mismatching page subjects. Single paid links (basically, placed manually) related to page topic have not been detected as spam. It speaks absence at links of spam features. The problem can be solved by the analysis of pages structure and revealing of advertising places, and also analysis of links life times (long pages monitoring is necessary).

References

[BCSU05] Benczur, A. A.; Csalogany, K.; Sarlos, T.; Uher, M.: Spamrank – fully automatic link spam detection. In *First International Workshop on Adversarial Information Retrieval on the Web*, Chiba, Japan, 2005; pp. 25–38.

[CDG+07] Castillo, C.; Donato, D.; Gionis, A.; Murdock, V.; Silvestri, F.: Know Your Neighbors: Web Spam Detection Using the Web Topology. *SIGIR'07*, May, 2007; pp. 423–30.

[Da00] Davison, B. D.: Recognizing nepotistic links on the web. In *AAAI–2000 Workshop on Artificial Intelligence for Web Search*, Austin, TX, 30 July 2000; pp. 23–28.

[FMN04] Fetterly, D.; Manasse, M.; Najork, M.: 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)*, Paris, France, 2004; pp. 1–6.

[EMT04] Eiron, N.; McCurley, K. S.; Tomlin, J. A.: Ranking the web frontier. In *Proceedings of the 13th International World Wide Web Conference (WWW)*, New York, USA, ACM Press, 2004; pp. 309–318.

[GGP04] Gyongyi, Z.; Garcia–Molina, H.; Pedersen, J.: Combating web spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, Toronto, Canada, Morgan Kaufmann, 2004; pp. 576–587.

[KR06] Krishnan, V.; Raj, R.: Web Spam Detection with Anti–Trust–Rank. In *the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '06)*, Seattle, USA, 2006; pp. 35–40.

[SS08] Sharapov, R.V.; Sharapova, E.V.: Link spam detection. In *Proceedings of the 10th RCDL Conference*, Dubna, Russia, 2008; pp. 191–196 [in russian].

[WD05] Wu, B.; Davison, B. D.: Identifying link farm pages. In *Proceedings of the 14th International World Wide Web Conference (WWW)*, Chiba, Japan, 2005; pp. 820–829.