Proceedings of the 7^{th} International Conference on Applied Informatics
Eger, Hungary, January 28–31, 2007. Vol. 2. pp. 417–422.

Search engine ranking

Mária Princz

Faculty of Technical Engineering, University of Debrecen
e-mail: pmaria@delfin.unideb.hu

Abstract

Ranking is a very important part of information retrieval systems. There are some ranking techniques, but search engine ranking algorithms are closely guarded secrets.

This paper deals with the mathematical basis of ranking, and details some principles that determine the relevancy of a web page. There are techniques to improve ranking. These techniques include two broad categories: techniques that search engines recommend as part of good design, and those techniques that search engines do not approve of because they try to manipulate search engine results.

Keywords: search engine, ranking, search engine optimization, spemming

1. Introduction

Search engine ranking algorithms are hidden for at least two reasons: Search engine companies want to protect their methods from their competitors, and they also want to make it difficult for web site owners to manipulate their rankings.

The paper is organized as follows. Section 2 gives a brief overview of mathematical basis of ranking. Section 3 discusses some view point of ranking. In section 4 we outline techniques to improve ranking. Finally, in Section 5 summary is given.

2. Mathematical basis of ranking

Traditional information retrieval systems usually adopt index terms to index and retrieve documents. The idea of this that the semantics of documents and of the user information need can be expressed through sets of index terms. Document relevance to the query dependents on a ranking algorithm. Relevance is implemented by the information retrieval model.

2.1. Classic models in information retrieval

There are three classic models in information retrieval: the Boolean, the vector, and the probabilistic models.

The *Boolean model* is based on set theory and Boolean algebra. Retrieval is based on whether or not the documents contain the query terms and makes return exact matches.

The traditional Boolean approach does not provide a relevance ranking of the retrieved documents, although modern Boolean approaches can make use of the location and frequency of keywords in document structure.

In the *vector model*, a document and a user query are represented as vectors in a t-dimensional space, where t equivalent with the number of index terms in the query. The index terms of the query are basis vectors in this space, and the document can express linear combination of the basis vectors. The coefficients equal 1, if the index terms are in the document and 0 otherwise. Document relevance to the query can be quantified by the cosine of the angle between these two vectors.

The *probabilistic model* ranks the documents based on the quotient of the probability that the document is relevant to query and the probability that the document is non-relevant to query. User relevance feedback is very important with this model.

3. How do search engines rank web sites?

3.1. Location and frequency of keywords

Most search engines use variations of the Boolean or vector model to do ranking, but the model does not matter too much. The ranking algorithms involve the location and frequency of keywords on a web page.

Search engines give special weight to keywords that appear:

- in the TITLE tag
- in the URL (such as in domain name, directories and file names)
- in HTML tags (headings, emphasized text)
- in other HTML tags (such keyword or description or ALT meta tag)
- in links pointing to the page
- keyword adjacency
- keyword proximity.

All the major search engines follow it to some degree but the location and frequency of keywords are easily influenced by webmasters. Therefore some search engine companies use other factors ("off the page" ranking factors) which are not easily influenced by webmasters. These factors are determined by web site popularity. Chief among these are link analysis (HITS algorithm, PageRank) and "click" popularity.

3.2. Link analysis

The number of hyperlinks that points to a page provides a measure of its popularity and quality. Not only is the quantity of links taken into account, but also the quality of the website linked to a page. Link popularity is used by every search engine to some extent.

3.2.1. HITS algorithm

It can be used to rank Web search results. Let S be the answer set. HITS uses two values for each page in set S, the authority and the hub value. Pages that have many links pointing to them in S are called authorities (they have relevant content). Pages that have many outgoing links are called hubs (they should point to similar content). The authority and hub value of page p are defined as follows:

$$a_p \leftarrow \sum_{q:(q,p)\in S} h_q, \quad h_p \leftarrow \sum_{q:(p,q)\in S} a_q.$$

3.2.2. PageRank

PageRank estimates the likelihood that a given page will be reached by a web user who randomly surfed the web, and followed links from one page to another. $T_1, \ldots, T_n$ pages point to page A with $PR(T_i)$ PageRank and $C(T_i)$ is defined as the number of links going out of page T_i. The parameter d is a damping factor witch is around 0.85. N is the total number of pages. The PageRank of a page A is given as follows:

$$PR_{j+1}(A) \leftarrow \frac{1-d}{N} + d\sum_{i=1}^{n} \frac{PR_j(T_i)}{C(T_i)}.$$

3.3. Click popularity

Click popularity is calculated by measuring the number of clicks each web site receives from a search engine's results page and record how long they stay at a website. It depends on how many users click on a link when it comes up on a search engine's results page and how long they stay at a website before returning to the ranked list.

Search engines Excite and Direct Hit use click popularity. We can increase click popularity by having a good title and a good description of our website in our meta tags.

4. Techniques to improve ranking

Market research indicates that 85% of search engine users do not go beyond the first page of results. Therefore the top positions in the result sets are very significant and valuable. Techniques to improve ranking include two broad categories:

- techniques that search engines are recommended as part of good design (SEO)

- and those techniques that search engines do not approve of because they try to manipulate search engine results (spamming).

4.1. Search Engine Optimization (SEO)

SEO is the process of improving a web site's internal elements (such as HTML titles, meta tags, etc.) and external factors (such as link popularity) to achieve high search engine rankings.

Search engine optimization as a subset of search engine marketing seeks to improve the number and quality of visitors to a web site. A SEO method is considered "White hat" if it conforms to the search engines' guidelines and/or involves no deception.

There are some major steps to successfully optimizing our web site:

Selecting the right keywords
This is the first and most important step. Study of keyword ranking, analysis and popularity – help to pick relevant keywords that drive relevant traffic to the site.

Optimizing the web site tags
HTML titles – the most important tag on the page.
Meta tags as keywords, description, ALT were very important some years ago. Webmasters tried to cram as many words as possible inside their meta tags. Nowadays, because of spam problems, meta tags are not even read by some search engines. (For example Google ignores the meta keywords tag.)
Content tags, such as header, bold, etc., are important.

Optimizing the content
Our keywords need to be reflected in the page's content.

Have HTML links
Search engine do not read image maps, so they can not follow these links. We solve this problem by adding some HTML hyperlinks to the home page that lead to major inside pages or sections of our web site.

Frames can kill
Some of the major search engines cannot follow frame links.

Dynamic doorblocks
Some of the search engines will not be able to index pages generating via CGI. Also, avoid symbols in our URLs, especially the ? symbol. Search engines tend to ignore it.

Submission
After the optimization process is complete, we will submit the website to all the major search engines and directories to raise link popularity. Submission is a term for search engine and directory registration. It means that they know our pages exist.

4.2. Spamming

It uses unethical means, "black hat SEO techniques" to unfairly increase the rank of sites in search engines. Search engines may penalize sites they discover using black hat methods, either by reducing their rankings or eliminating their listings from their databases altogether. There are some spamming techniques:

4.2.1. Content spam

Hidden or invisible text
When the text and the background color are the same, the font size is very tiny or it is possible to hide the text within the HTML code such as "no frame" sections, ALT attributes and "no script" sections.
Keyword stuffing
It means raising the keyword density or ratio of keywords to other words on the page.
Meta tag stuffing
Repeating keywords in the Meta tags, and using keywords that are unrelated to the site's content.
"Gateway" or doorway pages
These pages redirect visitors without their knowledge use some form of cloaking. Cloaking is a technique in which the content presented to the search engine spider is different from that presented to the users' browser.
Scraper sites
These types of websites are generally full of advertising, or redirect the user to other sites.

4.2.2. Link spam

Link farms
A link farm is any group of web pages that all hyperlink to every other page in the group.
Hidden links
Putting links where visitors will not see them in order to increase link popularity.
"Wiki spam"
Using the open editability of wiki systems to place links from the wiki site to the spam site.

4.2.3. Other types of spam

Mirror websites
Hosting of multiple websites all with the same content but using different URLs. To increase a site's ranking in a search engine by placing hyperlinks from each mirror to every other mirror (a technique known as link farming).
URL redirection
Taking the user to another page without his or her intervention. There are several

techniques to implement a redirect. The simplest one is the refresh meta tag. Some search engines give a higher rank to results where the keyword searched for appears in the URL.

5. Summary

In this paper, we have presented a general overview about search engine ranking. There are many factors involved in determining the rank order of search results.

This paper deals with the mathematical basis of ranking, and details some principles that determine the relevancy of a web page. In general, search engines use a combination of factors that always include keywords' location and frequency furthermore page popularity.

There are techniques to improve ranking. Search engine ranking algorithms are constantly being revised to improve performance and screen out 'webmaster tricks' that attempt to unfairly skew page ranking.

References

[1] Baeza-Yates, R., Ribeiro-Neto, B., Modern Information Retrieval, Addison Wesley, (1999).

[2] Brin, S., Page, L., The Anatomy of a Large-Scale Hypertextual Web Search Engine, WWW7 / Computer Networks 30(1-7), (1998), 107–117, http://www.stanford.edu

[3] Wikipedia http://en.wikipedia.org

[4] Monash, C., A helpful guide to Web Search Engines http://www.monash.com/

[5] SearchEngines http://www.searchengines.com/

[6] SearchEngineShowdown http://www.searchengineshowdown.com

[7] SearchEngineWatch http://searchenginewatch.com

Mária Princz
Faculty of Technical Engineering
University of Debrecen
4029 Debrecen
Ótemető u. 2–4.
Hungary