From Page Ranking to Topic Sensitive Page Ranking: Implementation and Impact

F.Rimbach[1,2], M.Dannenberg[2], U.Bleimann[2] and S.M.Furnell[1]

[1] Network Research Group, University of Plymouth, Plymouth, United Kingdom
[2] Institute of Applied Informatics Darmstadt (aida),
University of Applied Sciences Darmstadt, Germany
e-mail: frimbach@gmx.de

Abstract

The impact of changing search engine technology on small- and mid-sized Internet based companies is more significant than a variety of general micro- or macro-economic factors. Despite this, the implementation and impact of the underlying technology of search engines are understood by just a small amount of professionals. Explaining the calculation, implementation and impact of the PageRank and Topic Sensitive Page Ranking is the prerequisite to recapitulating existing search engine optimization strategies and to identifying new methods for leveraging the Internet for sales and marketing purposes. This paper aligns the complex calculations of the two concepts to enable a comparison and describes how different strategies have to be adapted to effectively attract potential customers.

Keywords

PageRank, Topic Sensitive Page Ranking, Internet Marketing, Search engine optimization

1. Introduction

Understanding the implementation and impact of the PageRank (PR) algorithm and the Topic Sensitive Page Ranking (TSPR) allows the focused development of Internet marketing strategies. This paper describes the calculation, implementation and impact of the two concepts and derives Internet marketing approaches based on search engine optimization (SEO). The objectives and contribution of this paper are the direct comparison of the technical implementation and marketing impact of the two algorithms and especially the illustration of feasible SEO strategies within both environments.

2. Page-Ranking

This chapter primarily introduces the calculation of the PR. Based on the calculation model, the implementation is illustrated. The descriptions of calculation and implementation are aligned to the TSPR concept to facilitate a comparison within the following chapter. Additionally, this chapter elaborates on the impact of the PR, different link strategies and feasible SEO approaches.

2.1 Calculation

The PR was developed by Page and Brin (Google, 2005) based on the BackRub search engine as well as works of Marchiori (1997) and Kleinberg (1999); the concept is patented under U.S. Patent 6,285,999 (United States Patent and Trademark Office, 2005a). The PR represents a

Web site's importance within a set of pages (e.g. Internet) and has a major impact on the positioning of Web sites within the search engine result pages (SERP).

Formula F_1 describes the rather simple calculation of the PR:

$$(F_1) \qquad PageRank(p_i) = \frac{q}{N} + (1-q) \sum_{p_j \in M(p_i)} \frac{PageRank(p_j)}{L(p_j)}$$

Within this formula q is the residual probability (usually 0.15) derived from the "random walk" principle (used to avoid rank sink), N the total number of pages, $M(p_i)$ the set of pages linking to p_i and $L(p_j)$ the number of outgoing hyperlinks of the page p_j. The PR values are mathematically speaking entries of the dominant eigenvector of the modified adjacency matrix underlying the Markov theory. Limiting the number of iterations required to efficiently calculate the precise PR is a major determinant to optimizing the speed of the crawling/indexation process. More information about an efficient PR calculation is illustrated by Ridings and Shishigin (2002). The overall ranking of a page within the SERP is deducted from the PR and the relevance-score (RS). As described in F_2 those two factors are weighted by a set of controls and a factor-base:

$$(F_2) \qquad Ranking(p_i) = [(1-d) + a(RS)] * [(1-e) + b(PR * fb)]$$

Within this formula, RS is the relevance-score (determined by onsite-factors like title-tag), PR the PageRank, as explained above, a, b, are weight controls and fb a factor-base to integrate the logarithmic core PR. While the PR is linear, the rank shown on the Google toolbar (http://toolbar.google.com) or alternative tools is mapped on a logarithmic scale with an approximate basis of 5-8. Considering this algorithm, it always requires a large increase of PR to achieve a higher rank within the Google toolbar.

PageRank	Logarithmic PageRank[5]
0.00000001 – 5	1
6 – 25	2
25 – 125	3
126 – 625	4
626 – 3,125	5
3,126 – 15,625	6
15,626 – 78,125	7
78,126 – 390,625	8
390,626 – 1,953,125	9
1,953,126 – infinity	10

Table 1: Mapping linear PageRank to logarithmic PageRank (Ridings and Shishigin, 2002)

As only the logarithmic PR is visible on the public Google toolbar, precise information about the real PR is not available. Additionally, the rank indicated on the Google toolbar is not always fully accurate. In some cases, the toolbar simply guesses the rank.

Considering the mathematical PR formula F_1, a closed system of one Web site with N Web pages (tree structure) can establish any page rank by maximizing N. To reach a rank of 10 on the logarithmic Google toolbar, approximately 4,300,000 Web pages have to be structured. As this number can be created using dynamic scripts ((un-)casting the dynamic hyperlinks as static via mod_rewrite (Apache, 2005)), the indexing algorithm has the functionality to

exclude specific sets of structures. (The time required for a full illustration of so many hyperlinks on one page of an optimal tree would easily time out the request.)
The PR perfectly represents the overall popularity of a Web page within the Internet and helps to identify high value search engine (SE) results. However, as the general PR has no connection to content specific information, it remains rather useless as long as no additional content specific factors are considered within the positioning of results on the SERP. This consideration is done by the RS, which measures the relevance of a Web site to a specific query based on a large set of indicators, e.g. keywords appearing in URL, title, meta-tags, headlines and body text of the Web page. From a SEO perspective, those elements are called onsite-factors. The concrete impact of the RS derived from these factors is explained below.

2.2 Implementation

Google was officially founded in September 1998; a year later, in September 1999, the system removed the beta-label (Google, 2005b). (The information below is based on Page and Brin (1998).) Google is fetching URLs coordinated by a URL-server. The fetched sites are zlib (RFC 1950) compressed and sent to a store-server indicated with a doc-id. An indexer distributes the parsed and analyzed Web page as hit lists (set of words) in a partially sorted forward index barrel. Additionally, the indexer stores information about hyperlinks on the document in an anchor file. Those hyperlinks, converted by a URL-resolver, are again associated with doc-ids. A database with pairs of such resolved doc-ids is used to compute the PR as stated above. Finally, the sorter creates an inverted index from the barrels sorted by word-ids. Based on this index, the program DumpLexicon builds a lexicon (sized to fit in the RAM of a computer), which is then used by the searcher.

Performing a single word search, Google converts the word into the word-id to search in the short barrel, analyzing the hit lists of the indexed documents. The hit type is combined with a type weight (the dot product of the vector of count-weights with the vector of type-weights) to the RS. For a specific query, Google uses onsite-factors to select a first subset of relevant matches (RM; example 10,000 pages) from the total number of matches (M; example 100,000 pages) from the large repository (approx. 27 billion pages). This subset RM is determined by approximately 2 simple indicators (presumably title-tag and keyword density (ratio of the number of occurrences of a particular keyword or phrase to the total number of words in a page)). The subset RM is then sorted applying the whole RS combined with the PR (F_2). From the sorted subset RM, the first 1,000 are shown on the SERP ordered by Ranking(p_i).

As explained, the onsite-factors are highly important for Web pages in such an algorithm; a high PR is totally insignificant in case the Web page does not fulfill the requirements for being included in RM. This non-PR threshold determines a set of different SEO strategies (Ridings and Shishigin, 2002).

2.3 Impact

The PR, which illustrates the popularity of a Web page, is a system which closely represents the searcher's desire to identify the major Web page to a specific subject. Understanding a high PR as a large set of qualitative votes from different Web pages M(p_i), the targeted Web page (p_i) is likely to be of high quality. Additionally, the utilization of a PR system as one key determinant in the ranking algorithm is a first step to increasing the difficulty of manipulating the SERP. A significant PR, one of the important factors for the overall ranking, can be built up mainly by continuously acquiring more inbound hyperlinks. As programs and simple

strategies facilitated to speed up the process of spreading hyperlinks, the whole idea of valuing inbound hyperlinks independently from any additional factors became disrupted.

The significance of PR for the overall ranking has stressed the necessity of distributing inbound hyperlinks within the Internet. Trying to spread inbound hyperlinks quickly lead to strategies like simple link-exchange and link-farms as explained below. This manipulation of references severely distorted the identification of valuable natural votes. Another weakness of the PR system is that it is continuously polarizing the popularity. A highly ranked Web page will naturally receive more and more inbound hyperlinks which promote the Web page with an even higher rank.

2.4 SEO-Strategies

Internet marketing strategies started to focus on SEO within the late 1990s. As SEO is always based on the technology of the specific search engine, SEO specialized on Google started at the end of 1999 when the company seriously emerged on the search engine market.

As stated within the previous descriptions, onsite-factors have been of primary importance to get into the subset RM. Onsite optimization, as the first SEO-strategy, mainly focused on the domain and URL selection, title-tag, meta-tags, header-tags, bold-tags and keyword-density. While the onsite optimization raised the RS over the threshold value, the PR had to be built up continuously. The establishment of more inbound hyperlinks has been approached by a) link-exchange and b) link spamming. While link-exchange was still a rather natural process ("I link to you, so you link to me"), link spamming was implemented in many different forms. Link-farms for example enabled the distribution of a large amount of static hyperlinks utilizing rather simple algorithms. More sophisticated technical approaches were the active utilization of the Web site's structure to highlight specific Web pages. The structures of the Web pages have an important impact on the PR distribution. Figure 1 describes 2x3 simple scenarios:

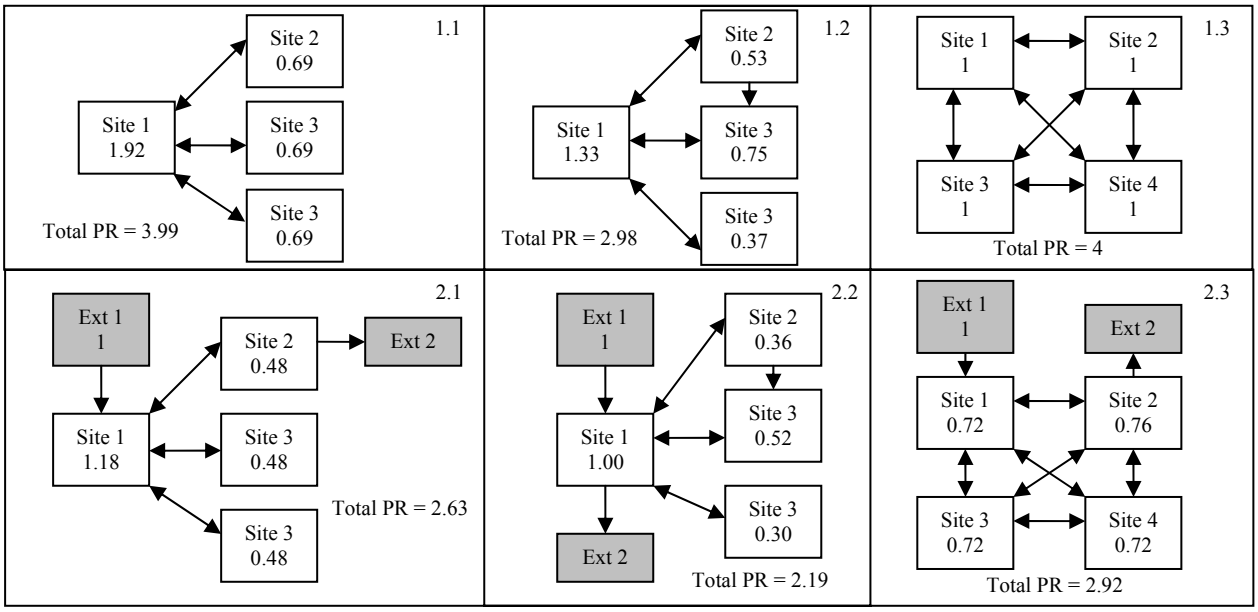

Figure 1 : Calculation of PR in Web page systems (values after 52 iterations)

Within scenarios 1.1-1.3 (closed system), it is clearly visible how to concentrate PR on specific Web pages in order to achieve a high ranking when competitive keywords (keywords

with a lot of search queries) are searched. While scenario 1.1 maximizes the PR of Site 1, scenario 1.3 distributes the PR equally on all pages. The natural site-structure of a tree is rather favorable. Scenario 1.2 creates a loss of PR. Within the open system (hyperlinks from and to external Web pages) of scenarios 2.1-2.3, it is visible that an outbound link from site 1 (in practice the home page) within the (common) tree structure causes a huge loss of total PR. As the scenarios 2.1-2.3 are closer to reality, the network structure (scenario 2.3) seems the most favorable.

This chapter explained that the PR implementation has an important impact on SEO strategies. Within the next part of this paper the TSPR concept and following SEO strategies are described.

3. Topic Sensitive Page Ranking

As implemented by Gerasoulis (2000) a ranking algorithm has to determine the subject-specific popularity (Teoma, 2005). Thies states correctly that the random walk principle, which is applied for the PR, is only applicable in case the Internet would cover a single subject (SEO Search Lab, 2005). Another method is the hypertext induced topic selection (HITS) as described by Kleinberg in the U.S. Patent 6112202 (United States Patent and Trademark Office, 2005b).

The concept of TSPR is documented by Haveliwala, a former Google employee and PhD student of Stanford University (Haveliwala, 2002). The TSPR adds a bias to the random walk theory by underlying a specific intent to the users walk within the Internet.

3.1 Calculation

With TSPR the indexing of the Internet is processed similarly to the classical approach. The difference in implementation of the TSPR starts with the identification of a set of base topics c_j. Haveliwala uses the 16 categories of the Open Directory Project (basically any alternative topic cluster can be used). In the process of indexing, all (16) TSPR vectors $c_j = PR(\alpha, v_j)$ (α = bias factor) are calculated for each Web page while $rank_{jd}$ is the rank of the document d for the topic j. In comparison to the classical PR, not a uniform but a non-uniform damping factor v_j is used. The following formula F_3, which describes the value of an inbound hyperlink, shows that an inbound hyperlink from a different topic cluster is not considered as a qualitative vote.

F_3

$$v_{ji} = \begin{cases} \frac{1}{|T_j|} & i \in T_j \\ 0 & i \notin T_j \end{cases}$$

For a specific query q, q_i' is the i^{th} term within the context of q. For each term in q' the value for the topic class c_j is evaluated (within the query-time) using the following formula F_4:

F_4

$$P(c_j, q') = \frac{P(c_j) \cdot P(q' | c_j)}{P(q')} \alpha P(c_j) \cdot \prod_i P(q_i' | c_j)$$

As shown in Table 2 below each query is indicated with a set of $P(c_j,q)$ values:

Bicycling	
Sports	0.52
Regional	0.13
Health	0.07

Amusement parks	
Regional	0.51
Recreation	0.23
Kids & Teens	0.08

Table 2: Estimated $P(c_j,q)$ for the queries "Bicyling" and "Amusement parks"

Equivalent to the previous concept, the calculation cannot be applied to all matches in the database. Therefore, the final algorithm is only used on a subset RM as described within the PR concept. The subset RM is then sorted by the query's topic sensitive importance score s_{qd} combined with a RS. The last formula F_5 illustrates the combination of the value explained above and the $rank_{jd}$ (the documents (d) rank vector for the topic c_j). Mainly just the topics j with the three highest values for $P(c_j,q)$ are required to calculate a precise s_{qd}.

F_5 $$s_{qd} = \sum_j P(c_j \mid q') \cdot rank_{jd}$$

3.2 Implementation

The implementation of a TSPR method in Google's algorithm is an assumption and not publicly communicated. Google acquired Applied Semantics in April 2003 (Google, 2005c). The CICRA technology of Applied Semantics is a scalable ontology system with a large database with words, their meaning and conceptual relation to other meanings. Additionally, CICRA facilitates the identification of how closely related two phrases are. It is obvious that this technology (as used in Google AdSense) has been used to fundamentally analyze the topic cluster of Web pages to calculate the TSPR.
Besides Google, the SE Teoma officially uses subject-specific popularity (Ask 2006), which they call Expert Rank.

On the 16th November 2003 the so-called "Florida Update" (a major change of Google's algorithms) and the implementation of the TSPR concept changed the SERP of Google massively. The differentiation of one general PR into a set of TSPR does not require a significant change of the other aspects mentioned within the implementation of the general PR.

An additional concept linked to the Florida Update was the so called Local-Score (LS).

F_6 $$Ranking(p_i) = [(1-d) + a(RS)] * [(1-e) + b(PR * fb)] * [(1- f) * c(LS)]$$

Within formula F_6, RS, PR, fb, a and b remain as stated above, c is an additional weight value, d, e, f are damping factors and LS a score computed from expert documents referring to the document p_i. While the term Local-Score is misleading with other developments within the SE technology, the concept of expert pages has to be discussed separately. The extraordinary significance of referring expert pages is mainly equivalent to an inbound link from a Web page with a specific high TSPR.

3.3 Impact

The impact of the modified ranking calculation can be easily described in two scenarios. Within the scenarios two different hyperlink structures (ceteris paribus for onsite-factors) are illustrated. While the Web site W_1 (scenario A) had inbound hyperlinks $M(p_{i1})$ from Web pages with different topics ($T>3$), Web site W_2 (scenario B) had inbound hyperlinks $M(p_{i2})$ from similar topics ($T<3$).

Within scenario A, W_1 had a PR determined by formula F_1. The PR was applicable for all concentrations of the onsite information. Under the new algorithm, W_1 suffered heavily as the inbound hyperlinks $M(p_{i1})$ did not focus on a specific topic. The total PR is literary distributed to all topic categories (T) and not concentrated on the Web pages' content. A mismatch of T_{W1} and the T of the Web pages with the outbound hyperlinks leads to fatal slip in the SERP.

Within scenario B, W_2 had a comparable general PR before the Florida Update. After the changes, the content-specific hyperlinks from $M(p_{i2})$ were concentrated on 2-3 specific topics. W_2 is ranking high within the SERP.

3.4 SEO-Strategies

Massa and Hayes (2005) state, that creating hyperlinks underlies an explicit intention by human intelligence. While it is true that active editors link more often to good content (positive hyperlinks), this first statement is false and can be utilized by SEO specialists. (In case of linking to distrusted sites (negative hyperlinks) the attribute "rel" can be used within the anchor-tag with a value "nofollow" (Technorati, 2005) – alternative approaches are described by the W3C (W3C, 2005). As more and more interfaces allow webmasters the (non-iframe) integration of offsite content on their Web pages, a feasible strategy within a TSPR environment is Trojan Link Distribution (TLD) (active and passive). While passive TLD optimizes a Web page W_1 to get integrated by another Web page W_2, active TLD spreads hyperlinks by self-requests or black-boarding in online communities or other open platforms. Passive TLD is good method to raise link popularity, active TLD can be considered as spamming.

Simple link exchange can be easily identified by the indexing algorithm of modern SE. The identification is based on the topic-comparison of W_1 and W_2 or the structure/position of the hyperlinks. A working link-exchange strategy now has to be closer to the natural process. Hyperlinks have to be surrounded by enriched content, placed on Web pages with favorable topic categories and finally referred from distributed network locations (IP). The Enriched, Categorized, Distributed Link Exchange (ECD-Link Exchange) is a simple but feasible extension of common link exchange programs.

4. Conclusion

This paper describes the implementation and impact of PR and TSPR. The gap of the PR, not being content specific, has been tackled by utilizing modern taxonomy technology. While the calculation of the primary rank value itself remains rather similar in both algorithms, the rank value for the TSPR algorithm is split into differentiated values for each topic. In a TSPR environment the SEO strategies require a further alignment to the natural process of setting qualitative references as votes out of a distributed network of Web pages.

Two major objectives coin the SE development: (1) identify the optimal information for a conducted search and (2) manage effectively and efficiently the SE index (while objective no. 1 partly determines objective no. 2). Even though PR as well as TSPR are automatically generated values which can be manipulated from the moment they are understood, an important and simple development can be identified: While SE more and more successfully utilize complex technology to identify relevant Web pages for a query, all SEO strategies head into the direction of building highly relevant Web pages. While the development of high quality Web pages is very positive for the user, the aspect of information and structure diversification is excluded. Available information tends to be homogenous; main positions of

the SERP basically illustrate equivalent content. Whether this homogeneity is favorable for the searcher has to be questioned. The future may ask for further approaches, especially for the presentation of search results.

References

Apache (2005), "Apache module mod_rewrite", http://httpd.apache.org/docs/1.3/mod/mod_rewrite.html, (Accessed 10th December 2005).

Ask (2006), "About Ask.com: Webmasters",http://sp.ask.com/en/docs/about/webmasters.shtml#how, (Accessed 3rd April 2006).

Brin, S., Page, L. (1998), "The Anatomy of a Large-Scale Hypertextual Web Search Engine", in Computer Networks 30(1-7), pp107-117.

Google (2005), "Google Technology", http://www.google.com/technology/, (Accessed 13th December 2005).

Google (2005b), "Google Corporate Information: Google Milestones", http://www.google.com/corporate/history.html, (Accessed 13th December 2005).

Google (2005c), "Google Press Center: Press Release", http://www.google.com/press/pressrel/applied.html, (Accessed 13th December 2005).

Haveliwala, T. H. (2002), "Topic sensitive Page Rank", in Proceedings of the 11th International Word Wide Web Conference, pp517– 526.

Kleinberg, J. (1999), "Authoritative sources in a hyperlinked environment", in Journal of ACM (JASM), 46.

Marchiori, M. (1997), "The quest for correct information on Web: Hyper search engines", in Proceedings of the 6th International World Wide Web Conference and Computer Networks and ISDN Systems, 29, pp1225-1235.

Massa, P., Hayes, C. (2005), "Page-reRank: using trusted links to re-rank authority", in IEEE/WIC/ACM International Conference on Web Intelligence 05.

Ridings, C., Shishigin, M. (2002), "PageRank Uncovered", http://www.voelspriet2.nl/PageRank.pdf, (Accessed 10th December 2005).

SEO Search Lab (2005), "Special Report: How to proper with a the new Google, by Dan Thies", http://www.seoresearchlabs.com/seo-research-labs-google-report.pdf, (Accessed 11th December 2005).

Technorati (2005), "Developers Wiki – RelNoFollow", http://developers.technorati.com/wiki/RelNoFollow, (Accessed 11th December 2005).

Teoma (2005), "Adding a New Dimension to Search: The Teoma Difference is Authority", http://sp.teoma.com/docs/teoma/about/searchwithauthority.html, (Accessed 11th December 2005).

United States Patent and Trademark Office (2005a), "Method for nod ranking in a linked database", http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=6,285,999, (Accessed 11th December 2005a).

United States Patent and Trademark Office (2005b), "Method and system for identifying authoritative information resources in an environment with content-based links between information resources", http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=6112202, (Accessed 9th December 2005b).

W3C (2005), "XML Linking Language (XLINK Version 1.0)", http://www.w3.org/TR/xlink/, (Accessed 11th December 2005).