Page ranking and topic-sensitive page ranking: micro-changes and macro-impact

F. Rimbach

Network Research Group, University of Plymouth, Plymouth, UK, and Institute of Applied Informatics Darmstadt (AIDA), University of Applied Sciences Darmstadt, Germany, and

M. Dannenberg and U. Bleimann

Institute of Applied Informatics Darmstadt (AIDA), University of Applied Sciences Darmstadt, Germany

Abstract

Purpose – The purpose of this paper is to examine the marketing and sales implications of page ranking techniques, in terms of how companies may use knowledge of their operation to increase the chances of attracting custom.

Design/methodology/approach – 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.

Findings – Different strategies have to be adapted to effectively attract potential customers.

Originality/value – This paper aligns the complex calculations of the two concepts to enable a comparison. The changing technology of search engines means that they are getting ever more complex – this article offers a snapshot of major developments.

Keywords Internet marketing, Search engines, Customers

Paper type Research paper

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. Certainly the changing technology of Search Engines (SE) means that they are getting ever more complex and this article is just a snapshot of major developments.

The paper begins with the illustration of the calculation, implementation and impact of PR in chapter 2. Further on SEO-strategies are described which were successfully used to leverage the algorithm to channel users to specific web sites. In chapter 3 the concept of TSPR is explained, again outlining calculation, implementation and impact. Finally 2 SEO-strategies, which can remain effective within this environment, are
Page ranking

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.

Calculation

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

Equation 1 describes the rather simple calculation of the PR:

\[
\text{PageRank}(p_i) = \frac{q}{N} + (1 - q) \sum_{p_j \in M(p_i)} \frac{\text{PageRank}(p_j)}{L(p_j)}
\]

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 Equation 2 those two factors are weighted by a set of controls and a factor-base:

\[
\text{Ranking}(p_i) = \left[ (1 - d) + a(\text{RS}) \right]^* \left[ (1 - e) + b(\text{PR}*f_b) \right]
\]

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 \( f_b \) 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 (see Table I).

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 \( PR \) Equation 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 \textit{PR} perfectly represents the overall popularity of a web page within the Internet and helps to identify high value 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 \textit{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 \textit{RS} derived from these factors is explained below.

\textit{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 Brin and Page, 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 \textit{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 \textit{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 (approximately 27 billion pages). This subset RM is determined by approximately two 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

\begin{table}[h]
\centering
\begin{tabular}{ll}
\hline
PageRank & Logarithmic PageRank_{5} \\
\hline
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 \\
\hline
\end{tabular}
\caption{Mapping linear PageRank to logarithmic PageRank (Ridings and Shishigin, 2002)}
\end{table}
The subset RM is then sorted applying the whole RS combined with the PR (Equation 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).

**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 manipulations, 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. In spite of the good algorithm, one of the biggest problems of Google was that many low quality web sites ranked high on the SERPs.

**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: link-exchange and 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 $2 \times 3$ 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. The scenario above focuses on the optimization via the internal hyperlink structure. Certainly the structures can be highly customized for more complex setups.

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.

**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 US 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.

**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 \( c_j = PR(\alpha, v_j) \) (\( \alpha = \) 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. Equation 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.

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

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) Equation 4:
\[
P(c_j, q') = \frac{P(c_j) \cdot P(q'|c_j)}{P(q')} \cdot \alpha P(c_j) \cdot \prod_i P(q'_i|c_j)
\]

As shown in Table II each query is indicated with a set of \( P(c_j, q) \) values.

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. Equation 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 \( s_{qd} \).

\[
s_{qd} = \sum_j P(c_j|q') \cdot rank_{jd}
\]

**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 16 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).

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

| Bicycling | 0.52 |
| Sports   | 0.13 |
| Health   | 0.07 |

Table II. Estimated \( P(c_j, q) \) for the queries “bicyling” and “amusement parks”
Within Equation 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.

**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 Equation 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 $TW_1$ 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 two specific topics. $W_2$ is ranking high within the SERP.

**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. One example of passive TLD is to leverage automatic inclusion scripts of other Web sites. Many inclusion scripts pull information from general indexes and it might be possible to rank high in this index. On the other hand active TLD means posting specific URLs in discussion forums (via texts or signatures), guest books or articles. Passive TLD is good method to raise link popularity, active TLD can be considered as spamming. The objective of introducing this strategy is not to motivate improper search engine marketing techniques but to simply raise the transparency about the weaknesses of automated evaluation algorithms. As mentioned further below, Google is continuously discussing the integration of human evaluation schemes into the overall search logic, but a consequent approach has not been implemented so far.

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.

Within the ECD concept, building enriched hyperlinks means to embed them into an environment of equivalent and descriptive content. It has to be evaluated in further studies whether a surrounding text of several words is sufficient for this approach or whether the whole web site’s content cluster determines the SE’s consideration as a valuable reference. While the first research shows that a short description of the hyperlink already has a large impact on the SE ranking, this does not comply with the existing knowledge about the strategies of SE implementation.

The aspect of categorization simulates the natural behavior of users. Identifying qualitative hyperlinks means to determine if the hyperlink is set to refer to more information about a specific topic. If web site A is within the content cluster “Arts” and links to another web site B of the content cluster “Arts”, it looks like a natural reference for the user who is interested in the topic “Arts” to look for more information about the topic on the web site B (scenario 1). Hyperlinks of the web site A to web site C, which is in the cluster “Health”, may not help the user to find more information about arts (scenario 2). The user potentially looks for information about arts; hyperlinks to other arts web sites are therefore interesting (scenario 1), hyperlinks to non-arts web sites are irrelevant (scenario 2). In scenario 2 the hyperlink from A to C is not considered as valuable and does not promote the ranking of the web site C or A.

The distribution of the incoming hyperlinks to external web pages is the final important aspect. External hyperlinks should be in a completely different hosting environment (IP-C or B block) and have different domain registration (WhoIs) details. Currently, hosts do not provide the service to distribute links on various locations. This might become a service offered by IM agencies.

**New developments**

From the end of December 2005 until March 2006 a major change of Google’s landscape called “Big Daddy” was implemented. The primary focus of the change was related to an update of Google’s datacenter infrastructure, also tackling canonical and 301 problems; the change was not intended to enhance any indexing or search algorithms or to perform a data refresh (Matt Cutts, 2006).

After “Big Daddy” the supplemental results were refreshed and SERPs changed significantly. Over the same period of 2005/2006 further quality rules were implemented. Web sites, which have outbound hyperlinks to low quality web sites, are penalized for these references. This step puts significant pressure on web masters not to spread bad hyperlinks. Even though this quality measure improves the overall PageRank system the above-mentioned strategies can still be applied.

A further discussion covers the integration of human evaluation schemes to the overall ranking calculation. While this approach would increase the indexation effort significantly a major improvement to a fully automated approach cannot be reliably predicted at this point of time. Leveraging the large group of Google users to perform this continuous evaluation could become a feasible alternative. One approach would be
to utilize traffic information (generated via the Google web site or the toolbar) to identify “good users” and rate their surf behavior as evaluations. Good users can be easily identified by their navigation behavior, especially if the profile is followed over a longer period of time. A positive rating for a web site can be derived from a repeated and long utilization. Certainly such automated scoring processes have to be proof against manipulation which can be assured by the proper identification of “good users”.

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.
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


**Corresponding author**

F. Rimbach can be contacted at: frimbach@gmx.de

---

To purchase reprints of this article please e-mail: reprints@emeraldinsight.com
Or visit our web site for further details: www.emeraldinsight.com/reprints