Online Information Review

Estimating Google's Search Engine Ranking Function from a Search Engine Optimization Perspective

Cheng-Jye Luh Sheng-An Yang Ting-Li Dean Huang

Article information:

To cite this document:
Cheng-Jye Luh Sheng-An Yang Ting-Li Dean Huang , (2016),"Estimating Google's Search Engine Ranking Function from a Search Engine Optimization Perspective", Online Information Review, Vol. 40 Iss 2 pp. -
Permanent link to this document:
http://dx.doi.org/10.1108/OIR-04-2015-0112

For Authors

About Emerald www.emeraldinsight.com

Estimating Google's Search Engine Ranking Function from a Search Engine Optimization Perspective

Abstract

Purpose –This study aims to estimate Google search engine's ranking function from a search engine optimization (SEO) perspective.

Design/methodology/approach – The paper proposed an estimation function that defines the query match score of a search result as the weighted sum of scores from a limited set of factors. The search results for a query are re-ranked according to the query match scores. The effectiveness was measured by comparing the new ranks with the original ranks of search results.

Findings – The proposed method achieved the best SEO effectiveness when using the top 20 search results for a query. Our empirical results reveal that PageRank is the dominant factor in Google ranking function. The title follows as the second most important, and the snippet and the URL have roughly equal importance with variations among queries.

Research limitations/implications –This study considered a limited set of ranking factors. The empirical results reveal that SEO effectiveness can be assessed by a simple estimation of ranking function even when the ranks of the new and original result sets are quite dissimilar.

Practical implications – The findings indicate that Web marketers should pay particular attention to a webpage's PageRank, and then place the keyword in the page title, snippet, and URL.

Originality/value – There have been ongoing concerns about how to formulate a simple strategy that can help a website get ranked higher in search engines. This study provides web marketers much needed empirical evidence about a simple way to foresee the ranking success of an SEO effort.

Keywords: Search Results Ranking, Latent Semantic Analysis, Search Engine Optimization,

Web Marketing

Article Classification: Research Paper

1. Introduction

Search engines such as Google and Yahoo have become the primary tools used to locate information on the internet. Several studies on user behavior indicated that most users click on websites listed on the first page of results and the proportion of users that view web sites listed beyond the third page of results decreases rapidly (Lorigo et al., 2006; Spink et al., 2006; Enge et al., 2012; Hopkins, 2012; Chuklin et al., 2013). As a result, achieving a high ranking in search engine results is crucial to attracting traffic to a website and represents the main thrust of search engine marketing efforts. However, achieving high ranking is exceedingly difficult because search engines do not publicly release their ranking algorithms nor disclose information regarding the factors used in ranking. Theoretical discussions of ranking algorithms (Langville & Meyer, 2006; Manning et al., 2009; Derhami et al., 2013) can help reveal the actual workings of search engines; however, they do not provide practical guidelines for webmasters or marketers to conduct search engine optimization (SEO).

SEO refers to the efforts intended to improve the ranking of a website in the search results for given target keywords (Gandour & Regolini, 2011; Moreno & Martinez, 2013; Berman & Katona, 2013). SEO technique includes two major processes: off-page optimization and on-page optimization. Off-page optimization entails building back links on other well reputed websites and thus boosting domain-level and page-level authority. On-page optimization SEO requires the optimization of web pages using target keywords in the title, snippets, and in the URL. The insertion of additional terms, semantically related to the target keyword, is considered an advanced SEO technique and is gaining popularity (Gennaro, 2015; Searchmetrics, 2015). Finding semantically related terms from small datasets is easy; however, this task can be troublesome for large datasets. This study sought to identify a set of terms semantically related to a given search query using latent semantic analysis of the search results retrieved from search engines (Google, in particular). A search query is a word or a set of words a user types into the search box. We call each word in a search query as a query term. Then the relevance scores of the terms semantically related to the search query provide the foundation for computing the query match score for a search result (or simply referred to as a document). For each document, the weighted sum of the query match scores related to the title, snippet, URL and off-page factors constitutes the query match score for the document. All of the documents retrieved for a given query are re-ranked according to the new match score. The newly generated ranks are then compared with their original counterparts to evaluate the effectiveness of the proposed method. This study differs with previous studies that re-rank search results with pseudo feedback relevance or external resources. Previous studies intend to improve the retrieval effectiveness so the users can find web pages most relevant to their query needs (Carpineto & Romano, 2012; Torkestani, 2012). This study instead focuses on how to measure SEO ranking success probability by a re-ranking approximation.

This study formulated the ranking function estimation problem as a curve-fitting process and constructed a simple mathematical function that best fits a series of documents for several queries. The factors of interest include details obtained directly from document's metadata, such as the title, snippets, and URL as well as publicly available Off-page factors such as PageRank, the page quality score by Google. We present the proposed ranking function as the weighted sum of these factors with near-optimal weights generated using a genetic algorithm.

This paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed method. Section 4 discusses the experimental results. Finally, Section 5 concludes this study and proposes future directions.

2. Related Work

2.1. *Search Ranking Factors and SEO*

Search engines rank search results according to a broad range of factors. Google is said to employ more than 200 factors in its ranking algorithm; most of which Google held as closely-guarded secrets. According to the SEO starter guide (Google 2010), the factors related to search ranking include the title, meta-description, anchor text, and various other on-page content-based factors. However, this guide barely mentions off-page (query-independent) factors, such as PageRank and the number of external links. Cutts (2011) conceptualizes the 200 plus factors into two categories: Trust - an assessment of a site's authority and reputation, and Relevance - an assessment of how well a page or site relates to a specific user query.

No SEO firms or industry professionals know the ranking details of any search engine. Even very successful SEO tools companies like Sistrix[1] have difficulty figuring out the extent of change or what factors were weighted differently after a major Google algorithm update. Previous efforts to decipher the process of ranking details involve experimentation and observation. For example, Evan (2007), after examining fifty specially optimized web pages, discovered that the seven most popular techniques used by SEO practitioners are not necessarily effective. Of particular interest was that websites with a high PageRank have no guarantee of obtaining high ranking. Zhu & Wu (2011) derived top five factors for SEO including URL length, Keyword in URL domain, Keyword density in H1, Keyword density in title, and URL layers based on a simple analysis of web pages content. Moreno & Martinez (2013) claimed that conducting search engine optimization for websites to gain high visibility in search engine results also makes their content more accessible. Sagot et al. (2014) suggest that SEO process should use an integrated meta-model to cope with the variability and

[1] http://www.sistrix.de/google-updates/bahn.de/

fluctuation of webpages ranking. The models employed include a conceptual model that transfers user needs into SEO objectives, an experimental model that implements the SEO practice to fulfill SEO objectives and a computation model that monitors the ranking of target web sites.

Recently, several search metrics and SEO tool companies regularly published their ranking factors study reports. For example, SEOmoz[2] publishes a ranking factors report bi-annually based on survey of search professionals and data analysis results of the top 50 search results for over 10,000 queries across multiple categories. Similarly, Searchmetrics[3] has published an annual ranking factors study since 2012. Both firms unanimously claimed that the factors used in these reports are not what are being used in Google's ranking algorithm, but simply show the features of web pages that tend to rank higher. These studies generally use rank correlation coefficient to indicate the relationship between the rankings of search results and the feature values (e.g., total number of external links) of the search results on a per-feature basis. The factors that have relatively high correlation coefficient are considered to have strong influence on search engine ranking. Notably, the so-called factors are examined one by one, but not in groups of any combination form, for their individual correlation with search results rankings.

According to the ranking factor survey performed by SEOmoz (2011, 2013), the top three most important factors categories are domain-level authority link metrics, page-level link metrics, and page-level content based metrics. Domain-level authority link metrics includes the number of root domains linking to the domain, the number of unique IPs linking to the domain, mozRank of the domain, and many more. Page-level link metrics entail link metrics to the individual ranking page such as number of links and MozRank. Page-level content based metrics describe the use of the keyword in HMTL code of web page such as the title

[2] https://moz.com/search-ranking-factors/survey
[3] http://www.searchmetrics.com/

tag, the body, the meta description, and the H1 tags. SEOmoz conclusively suggests that links are still the most important part of the algorithm and keyword usage on the page is still fundamental. Mavridis & Symeonidis (2015) presented a spearman correlation coefficient analysis of all the critical ranking factors based on data from three well known web metrics sites. The authors argue MozRank, Moz Page Authority, and Moz Domain Authority are highly correlated with search engine rankings. Additionally, the number of total links indicates positive correlation with search engine rankings. The ranking factor report by Searchmetrics (2015) reveals that backlinks are still an important ranking factor. Meanwhile, the relevance of keywords in the description is falling, but the percentage of topic relevant terms on high ranking websites has increased yet further.

The estimation of search ranking takes into accounts both content-based and query-independent features. PageRank, Google's static page score, is the commonly used query-independent ranking factor (Agichtein et al., 2006; Bifet et al., 2005; Fortunato et al., 2008; Richardson et al., 2006; Hariri, 2011). Unfortunately, PageRank is updated once every few months, and the PageRank results that are made publically available differ from what Google uses in their ranking algorithm (Bifet et al., 2005). Previously, SEO practitioners relied on the Yahoo Site Explorer for free access to valuable information, including external links and domains linked to websites; however, this service was shut down in November 2011. An alternative measure, MozRank, has attracted considerable attention. MozRank[4] is a link popularity score indicating the static importance of any webpage on the internet. Similar to PageRank, pages earn MozRank via the number and authority of other pages that link to them. To the best of our knowledge, it is the only publicly available query-independent score comparable to Google's PageRank. This study considers either PageRank or MozRank as a comprehensive ranking factor to represent the inclusion of domain-level and page-level authority link metrics.

[4] http://www.seomoz.org/learn-seo/mozrank

2.2 Effectiveness of Re-ranking

The effectiveness of search results re-ranking can be assessed from different perspectives. For example, SEO practitioners are interested in getting their target web pages ranked higher in search results. Information retrieval (IR) researchers are interested in the ways that would further improve the ranking of results returned by search engines.

From an SEO perspective, the effectiveness of re-ranking search results by a simple approximation require measuring how well the proposed method could still rank target web pages within top search engine results pages. The focus is on how many top ranked web pages remain listed there after the re-ranking approximation.

From an IR perspective, the purpose of re-ranking search results is to measure how well a proposed ranking method can further improve the ranking of results returned by search engines. Query expansion via relevance feedback is one of the major approaches often taken by IR studies to enhance retrieval effectiveness (Manning et al., 2009; Carpineto & Romano, 2012; Torkestani, 2012). Relevance feedback requires the user to give feedback on the relevance of documents in an initial set of outcomes. Then based on the user feedback the system expands or reformulates query terms needed for a revised set of outcomes. Pseudo relevance feedback instead uses the top-ranked documents in the initial retrieval as relevant feedback and requires no user input at all. Additionally, external resources, e.g. Wikipedia, have been widely employed to augment the initial search in pseudo-relevance feedback. Several studies have reported significant improvement in retrieval performance on this subject (Masri et al. 2013, Al-Shboul & Myaeng, 2013).

2.3 Latent Semantic Analysis

Latent Semantic Analysis (LSA), also known as Latent Semantic Indexing (LSI), is a document analysis method used to uncover relationships among documents and the terms they contain as well as the semantic relationships among terms (Baeza-Yates & Ribeiro-Neto, 2011; Minhas & Kumar, 2013). For the purpose of ranking, the semantic relationship between a term and a given query is evaluated as a relevance score. So each document can be scored on the base of the relevance scores of the terms it contains.

The input of LSA is a term-by-document matrix *A*. The rows in *A* represent terms, and the columns represent documents. Matrix *A* contains non-zero and zero elements. A non-zero element typically indicates the *tf-idf* value of a given term in the corresponding document, while a zero element denotes that a particular term does not appear in the corresponding document (Baeza-Yates & Ribeiro-Neto 2011).

The critical step in LSA is the factorization of matrix *A* using singular value decomposition (SVD) (Baeza-Yates & Ribeiro-Neto, 2011; Mirzal, 2013). The SVD of an $m \times n$ matrix *A* is given by USV^T where *U* is an $m \times r$ orthogonal matrix, S is an $r \times r$ diagonal matrix with nonnegative real numbers on the diagonal, and V^T is an $r \times n$ orthogonal matrix. Intuitively, *U* and *V* can be considered as the term-to-concept and document-to-concept matrix, respectively. The diagonal entries of S are the singular values of *A*. Then a dimensionality reduction process is conducted to find the most significant *k* singular values while dropping out the rest singular values. The most significant *k* singular values intuitively indicate the most important hidden concepts or dimensions in the term-document matrix. Maintaining the largest *k* singular values is equivalent to preserving the first *k* columns of $U$and the first *k* rows of S and V^T. Finally, multiplying the reduced versions of *U, S,* and V^T generates the reduced version of *A,* such that $A_k = U_k S_k V_k^T$ as shown in Fig. 1. A_k reveals the hidden semantic structure in the terms and relevant documents.

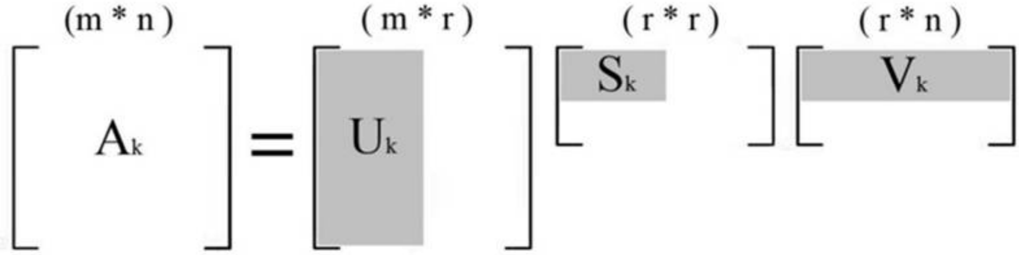
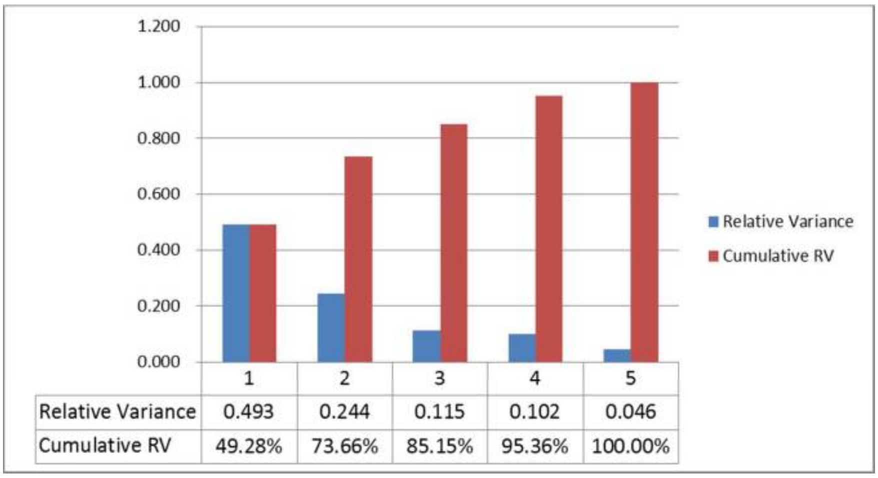


Figure 1. Singular Vector Decomposition

Creating a histogram of the relative variance of each singular value from the S matrix is the technique used to select the most significant *k* dimensions. The relative variance of a singular value s_j is defined as the square of s_j divided by the sum of the square of all singular values (Wall et al., 2003). The value of *k* is the cutoff point where the cumulative relative variance of some leading dimensions exceeds a predefined dimension reduction threshold. Given an example of S matrix with singular values {0.29, 0.204, 0.132, 0.113, 0.089 …}, its histogram of cumulative relative variance is shown in Fig. 2. We can chose k = 3, that is to select the first three dimensions, if the dimension reduction threshold is set at 0.8.



	1	2	3	4	5
Relative Variance	0.493	0.244	0.115	0.102	0.046
Cumulative RV	49.28%	73.66%	85.15%	95.36%	100.00%

Figure 2. Example of cumulative relative variance

3. Proposed Method

The proposed method includes three major steps: data crawling and preprocessing, latent semantic analysis, and document scoring and re-ranking as shown in Fig. 3.

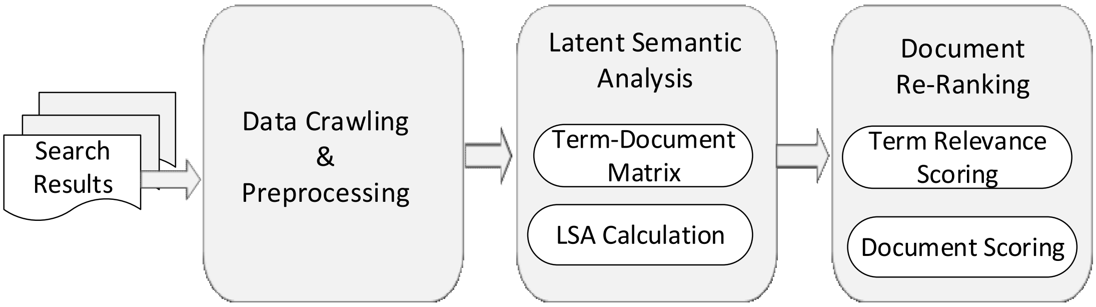


Figure 3. System Flow Chart

3.1. Data Crawling and Preprocessing

This step involves the retrieval of search results for a given query as well as data preprocessing. Search engines typically present search results as a ranked list of documents (or web pages). Once retrieving search results from a search engine, we conduct data preprocessing to parse out the title, snippet, and URL of each document as well as multi-word phrases from the parsed titles and snippets. The multi-word phrases are determined using a pairwise mutual information based n-gram package developed by Kazembe (2011).

The reason we adopt multi-word phrases as semantically related terms is that Google is believed to employ phase-based indexing and searching in their search engine, as revealed in a major Google patent (Patterson, 2009). For simplicity, both single words and multi-word phrases are referred to as "terms" in this study.

3.2 Latent Semantic Analysis

The first step in the latent semantic analysis is the construction of a term-document matrix. The rows of the matrix comprise the terms obtained from data preprocessing, and the columns represent the documents. Each entry in the matrix is a term's *tf-idf* score, indicating the significance of a term to the corresponding document in which the term appears. We used the open source Java Matrix Package, JAMA[5] to conduct the singular vector decomposition, that is, to transpose the original term-document matrix A into a dimension reduced version, such that $A_k = U_k S_k V_k^T$ as shown in Fig. 1.

During the SVD process, the dimension reduction technique presented in Section 2.3 is applied to select the k most important dimensions. We examined the impact of the dimension reduction threshold on re-ranking effectiveness and presented the results in Section 4.

3.3. Document Scoring and Re-Ranking

This step calculates a query-document match score for each document and then arranges the documents in descending order of this score. Thus, three things require to be accomplished. First, the relevance score of each term with respect to the given query must be determined by the reduced term-to-dcument matrix. Second, the query-document match score for each document is determined by combining the relevance scores of the terms it contains and the scores of other ranking factors, such as URL. Finally, the documents are re-ranked according to the resulting query-document match scores.

3.3.1. Determining the Relevance Scores of Term to Queries

This study quantified a term's relevance score to a given query by computing the cosine similarity of their vector representations obtained from the corresponding rows in the reduced U matrix, i.e., the term-to-document matrix after LSA process. For example, "chemical", "ecosystem", and "restoration" are some of the terms relevant to the query *oil spill.* Their relevance score to *oil spill* are 0.5409, 0.3531, and 8951 respectively from our experimental results.

[5] http://math.nist.gov/javanumerics/jama/

3.3.2. Document Scoring

A query-document match score comprises components from on-page factors and off-page factors, as suggested in literature (Bifet et al., 2005; Burges et al., 2005; Fortunato et al., 2008; Richardson et al., 2006; Su et al., 2010; Hariri, 2011; SEOmoz, 2011; 2013). The on-page factors considered in this study include a web page's title, meta description, and URL. On the other hand, we chose PageRank and/or MozRank as the only off-page factor. The importance of title, meta description, URL and PageRank to ranking are well known as indicated by high correlations in several ranking factors studies (Searchmetrics, 2015; SEOmoz, 2011; 2013;). This study does not intend to identify their importance to search ranking. Instead, we seek to determine their individual weights attributed to the query-document match score if they are used as ranking factors. The reason we only use title, meta description and URL as on-page factors is that they are simply accessible from Google search results pages and are visible to search users. The minor difference is that meta description is called snippet in search engine terminology. Similarly, both PageRank and MozRank are accessible via publicly available application programming interfaces. Notably, both PageRank and MozRank are defined to include both link popularity and authority components. Using either PageRank or MozRank as a comprehensive off-page factor can avoid the efforts needed to fetch the backlinks and other signals related to links ahd authority for every search results under study. This shortcut simplifies the tedious data access work, especially for SEO practitioners without huge computation and bandwidth support.

This study defines the query-document match score of document *d* with respect to query *q* as the weighted sum of title score S_T, snippet score S_S, URL score S_U, and PageRank/MozRank score S_R. Moreover, the weights of factors, namely, W_T, W_S, W_U, and W_R, are determined using a genetic algorithm.

$$Score(d,q) = W_T \cdot S_T(d,q) + W_S \cdot S_S(d,q) + W_U \cdot S_U(d,q) + W_R \cdot S_R(d) \qquad (1)$$

The query match scores for the title, snippet and URL are calculated as follows. Notably, the calculation of title-query and snippet-query match scores consider both query and non-query terms identified with SVD, but the calculation of URL-query scores only takes query terms into account.

1. Title-query Match score

Consider query *q* with *k* query terms *t_1, t_2,…, t_k*. The title-query match score, as denoted $S_T(d,q)$ in equation (1), is defined as follows:

Title-query match score = Query terms in title score + Non-query terms in title score (2)

where the first component counts the query terms in title score, and the second one is the sum of the relevance scores of non-query terms that appear in the title. We then define the query terms in title score by the following:

*Query terms in title score = Query proximity * Query prominence in title* (3)

where *Query proximity* computes how close together the query terms are in the title and *Query prominence* measures how close of the query terms are to the beginning of the title. The design rationale is: the closer of the query terms the higher query proximity, and the close of the query terms to the beginning of the higher query prominence.

Query proximity considers only the query terms that appear in order as a sub-phrase of the query. Consider the query *"liquid packaging machine"* as a concrete example: the set of its sub-phrases would be {*liquid packaging machine, liquid packaging, liquid machine, packaging machine, liquid, packaging, machine*}. For each member of the sub-phrase set, the sum of the reciprocal rank of the terms it contains is calculated as its original sub-phrase score. Then a normalized weight is obtained by dividing its original score with that of the longest sub-phrase, i.e., the query itself. For example, the normalized sub-phrase weights for *liquid packaging machine, liquid packaging*, and *packaging machine* are (1/1+1/2+1/3)/(1/1+1/2+1/3)=1, (1/1+1/2)/(1/1+1/2+1/3)=0.82, and (1/2+1/3)/(1/+1/2+1/3)= 0.45, respectively. Once more than one sub-phrases of the query appear in a given title, (e.g., ***Liquid*** *plastic* ***packaging machine*** *from* ***packaging machine*** *expert*), we select the one with the highest normalized weight (i.e., *liquid packaging machine*) as the matching sub-phrase to the given query. Notably in this example, the matching sub-phase is loosely coupled, that is, each member follows its order in the query, but may not exactly occur next to each other. The query proximity comes into measuring the closeness of the query terms in the title as follows:

Query proximity

= Word count of the matching sub-phase / Word count of the smallest window in title containing all terms of the matching sub-phrase (4)

We then define *query prominence* as the average of the prominence values of all query terms that appear in the document title. The term prominence in a title is given by the following:

Term prominence= (Title length – average offset to its position in query)/Title length (5)

By this definition, a multiply occurred query term increases the average offset to its original position in the query and has a lower term prominence. This design can prevent keyword spamming from obtaining a good prominence.

Consider the query *oil spill* and examples in Table 1. For the title "**Oil spill** - Wikipedia, the free encyclopedia", both the query promixmity and the query prominence have a value of 1 so its query-match score is 1. The reason is that both oil and spill appear at the beginning of the title and are in exactly the same order as in the query. For the title "**Oil Spill** Response - **Oil Spill"** the query term *oil* appears twice at the first and fourth positions of the title, its average term offset is ((4-1) + (1-1))/2= 1.5. Thus, the term prominence of *oil* is (5 - 1.5)/5=

0.7. Similarly, the term prominence of *spill* is 0.7. Then the query prominence for *oil spill* is 0.7. For the last title "**Oil** and Chemical **Spills**: NOAA Watch", the smallest window size including all query terms is 4 and the word count of the matching sub-phrase (i.e., **Oil Spills**) is 2. Thus, its *query proximity* is 0.5.Then the term prominences for *Oil* and *Spill* are (6-0)/6 =1 and (6-2)/6= 0.67, respectively. Then the *query prominence* is (1+0.67)/2 = 0.84. We believe the ranking of the title-query match score is very consistent with what users perceive.

Table 1 Title-query Match Score Examples for search query "Oil Spill"

Title	Query Proximity	Query Prominence	Match Score
Oil spill - Wikipedia, the free encyclopedia	1	1	1
Oil Spill Response - **Oil Spill**	1	0.7	0.7
Oil and Chemical **Spills**: NOAA Watch	0.5	0.84	0.42

For the last example, one could actually decompose the phrase into "Oil spills" and "chemical spills". From a query matching perspective, the title "Oil and Chemical Spills: NOAA Watch" has a lower query proximity, but a higher query prominence for query "*Oil Spill*" than that for query "*Chemical Spill"* as shown in Table 2. Thus, this document title has a higher match score to the query "*Chemical Spill*" than to the "*Oil Spill.*" Notably, the proposed method is query oriented, that is, we find "Oil * * Spills" and "Chemical Spills" as the matching sub-phase to query "Oil Spill" and "Chemical Spill" respectively. We do not interpret the title has the phrase "Chemical Spills" as the dominant matching sub-phase for all queries.

Table 2 Title-query Match Score Comparisons for “Oil and Chemical Spills: NOAA Watch”

Query	Query Proximity	Query Prominence	Match Score
Oil Spill	0.5	0.84	0.42
Chemical Spill	1	0.75	0.75

2. Snippet-query Match Score

The snippet-query match score, $S_S(d,q)$, is defined as the sum of the relevance scores of all terms occurring in the snippet of document *d* with respect to the search query *q*. Notably, each term is counted only once regardless of how many times it occurs in the snippet.

3. URL-query Match Score

We calculate the URL-query match score based on the following URL syntax:

http:// <domain>/[<path>]/[<filename>].

That is, a URL contains one or more labels separated by the slash symbol “/”. Among the labels, <domain> is assumed to have a canonical form, such as *www.example.com* or *example.com*, and both <path> and <filename> are optional. Within the domain name *www.example.com*, *.com* is one of the predefined top-level domains (TLD), *example* is the second-level domain of the .com TLD, and *www* is the host name. Query terms may appear in the second-level domain and/or the host name of a domain name. Intuitively, a URL with only one <domain> label with all query terms appearing in this part is preferred. On the contrary, a URL with query terms scattered throughout more than one label is unfavorable concerning query match scoring.

The URL-query match score is defined as the average of all URL label match scores. The computation of each URL label match score considers the rank of the URL label in which query terms appear and whether all of the query terms appear in order. We denote these two components as URL label weight and query proximity in URL as follows:

$$URL\ Label\ Match\ Score = URL\ Label\ Weight * Query\ Proximity\ in\ URL \quad (6)$$

The *URL label weight* for the i^{th} label of a URL is defined as the reciprocal of *i,* if some query terms appear in this label. That is, the URL label weights for <domain>, <path> and <filename> labels are 1, 1/2, and 1/3, respectively. Additionally, we introduce a penalty for URL in a non-canonical form. For example**,** the URL label weight of *www.instant**oilspill**.com* for the query *Oil Spill* is 1 and the URL label weight for ***home****.instant**oilspill**.com* is $1*(1-p_1)$. The value of p_1 can be adjusted as desired and it was 0.2 during the experiment.

Query proximity in the URL considers the matching of sub-phrases related to the query in the dominant URL label. It is defined similarly as that of query proximity in the title, except that here we count the character length of the string concatenation of the search query and the matching sub-phrase in URL. Similarly, we introduce a second penalty for the extra characters leading or trailing the matching sub-phrase. For example, the query proximity in URL for *www.instant**oilspill**.com* to the query *Oil Spill* is $(8/8)*(1 - p_2)$. The reason is that the character sequence *oil spill* appears in the domain part, and some leading characters (*instant)* exist ahead of *oil*. Here string preprocessing is required before comparing the query phrase string with the URL string. In the above example, the query phrase *oil spill* is concatenated into the character sequence *oilspill,* and URL strings with a hyphen or underscore such as *oil-spill* and *oil_spill* are normalized with the same character sequence *oilspill*. The value of p_2 can be adjusted as desired and we set it 0.2 in the experiments.

3.3.3. *Document Re-Ranking*

The query-document match score is the sum of the weighted query match scores of a document's title, snippet, URL, and PageRank or MozRank score according to equation 1. The weights of factors are determined using a genetic algorithm. We encode the scores of factors as an array and normalize each array element to the range of 0 and 1 to facilitate genetic encoding. The initial population includes genetic representations of documents for all of the queries being considered. During each successive generation, a query-document match score is computed for each document according to the obtained weights, and the documents returned for each query are then re-ranked in the decreasing order of the resulting scores. Thus, each document obtains a new rank with respect to a given query at the end of each generation. For example, based on the weights we obtain after a genetic run, Table 3 shows the new doc scores we calculate based on the equation (1) for some sample documents. After re-ranking by the new doc scores, the re-ranked order becomes [d1, d3, d2].

Table 3 Document Scoring Examples.

Document	Title Score	Snippet Score	URL Score	PageRank Score	Doc Score	New Rank
Oil spill - Wikipedia, the free encyclopedia https://en.wikipedia.org/wiki/**Oil_spill**	1	0.56	0.5	0.6	0.586	1
Oil and Chemical **Spills**: NOAA Watch http://oceanservice.noaa.gov/hazards/**spills**/	0.42	0.84	0.31	0.6	0.499	3
Oil Spill Response - **Oil Spill** http://**oilspill**responseproject.org	0.7	0.72	0.8	0.4	0.548	2

*Based on the weights of factors: $W_T = 0.048$, $W_S = 0.01$, $W_U = 0.325$, $W_R = 0.617$

*PageRank score is normalized to the range of [0..1] by dividing the PageRank by 10.

The objective of the genetic algorithm is to identify a set of weights that minimize the sum of squared rank error (SSRE) of the new rank and its original rank of each document being evaluated. The minimization of SSRE indicates pushing the new rank of each document close to its corresponding original rank, that is the re-ranked order should be as close to the original rank order as possible. During experimentation, the genetic process is repeated until either reaching the condition the change of SSRE in the last 100 trials is less than 0.01% or having consumed a pre-defined number of runs.

4. Experiments and Evaluation

4.1. Experimental setup

To evaluate the effectiveness of the proposed method, we chose one hundred keywords related to industrial products as search queries for collecting experimental data. These keywords are from various categories, including electronics, machinery, plastics, and LEDs. Each keyword was submitted as a search query to Google's International English site (www.google.com/ncr) to retrieve the top *n* search results for effectiveness evaluation.

The objectives of the experiments were twofold. First, we determined the values of the key parameters, including the number of top search results should be retrieved and the dimension reduction threshold for selecting critical dimensions in the construction of SVD reduced document-term matrix to enhance re-ranking effectiveness. Second, we identified the combination of factors that delivers best performance in estimating the Google ranking function.

To fulfill the first objective, we employed ranking factors including title, snippet, URL, and PageRank in document scoring of various numbers of documents ranging from 20 and 100 in increments of 10. For each increment, we evaluated six threshold values for dimension reduction: 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0. The case of 10 documents was excluded from the evaluation because using the top ten documents only is insufficient to perceive semantic differences between documents listed on the first page of results and those on the remaining pages. Additionally, we ignored the dimension reduction threshold values under 0.5 because the most important dimensions of interest should contribute more than 50% of the relative variance accumulated by all dimensions.

Then four cases involving various combinations of factors were empirically evaluated and compared on the base of the key parameter values found. The first instance considered only on-page factors: title, snippet, and URL. The second case added PageRank. The third case replaced PageRank with MozRank. Finally, the fourth case combined all five factors altogether. The four cases were denoted as T+S+U, T+S+U+PR, T+S+U+MR, and T+S+U+PR+MR, respectively. The additional off-page factor, i.e. MozRank, was employed in the experiment because Google makes updates to PageRank on an irregular basis and the PageRank values publicly available via Google API do not accurately reflect the current statuses of web pages.

4.2.Effectiveness metrics

This study employs Kendall's rank correlation coefficient (Abdi, 2007) and R-Precision (Manning et al., 2009) to evaluate the effectiveness of the proposed method as shown in Table 4. The Kendall's rank correlation coefficient, commonly referred to as Kendall's Tau coefficient, is based on counting the number of pairwise disagreements between two ordered lists. We used it here to measure the degree of similarity between the original and new ranks assigned to the same set of search results. R-Precision is originally used to measure the precision of the top R documents retrieved in a ranked information retrieval system. The calculation of R-Precision requires having a set of known relevant documents and counts how many known relevant documents are ranked within the top R positions. This study considers the top R documents in Google's search results to a given query as the pseudo set of known relevant documents. Then the R-Precision calculates the ratio of the number of correct hits to the number of R. Here a correct hit, from an SEO perspective, means a document ranked within top R in Google's search results is still listed well within top R positions after the proposed re-ranking and does not imply this document is truly relevant to user interest. We calculate an R-precision value for each query and average the R-precision values for *m* queries in the experiment as the final R-Precision. The value of R can be adjusted according to the number of top positions of interest. Typically, R = 10 is considered a good SEO practice.

Table 4 The Metrics used in this Study

Metrics	Purpose
Kendall Tau Coefficient	To measure the degree of similarity of the original and new rankings of search results
R-Precision	To measure how many of top R search results are still ranked within top R positions after being re-ranked with the proposed method

4.3.Experimental Results

Fig. 4 and Fig. 5 show the effects of dimension reduction threshold values on R-precision and Kendall's tau coefficient according to dataset size. For both measures, the dataset size of 20 documents outperformed the other dataset sizes across all dimension reduction threshold values while the other dataset sizes show mixed results. Analysis of variance demonstrates that the effect of dataset size was significant for R-Precision (*F(8,45)=81.09, p<0.01*), and for Kendall's tau coefficient (*F(8,45)=111.02, p<0.01*). Furthermore, a paired t-test indicates that the proposed method performed significantly better with a dataset of 20 documents than with other dataset sizes, in both R-Precision and the Kendall's tau coefficient.

Fig. 6 shows the mean R-precision and Kendall's tau coefficient for the four combinations of factors. Among them, T+S+U+PR+MR scored the highest, achieving a Kendall's tau coefficient of 0.154 and R-Precision of 0.59; T+S+U received the lowest scores with a

Kendall's tau coefficient of 0.124 and R-Precision of 0.574. Obviously, adding off-page factors, either PageRank (PR) or MozRank (MR) or both, to on-page factors improves both measures. The results also reveal that MozRank is comparable to PageRank in estimating the Google ranking function using the proposed method while T+S+U+MR outperformed T+S+U+PR in R-Precision, but T+S+U+PR surpassed T+S+U+MR in Kendall's tau coefficient.

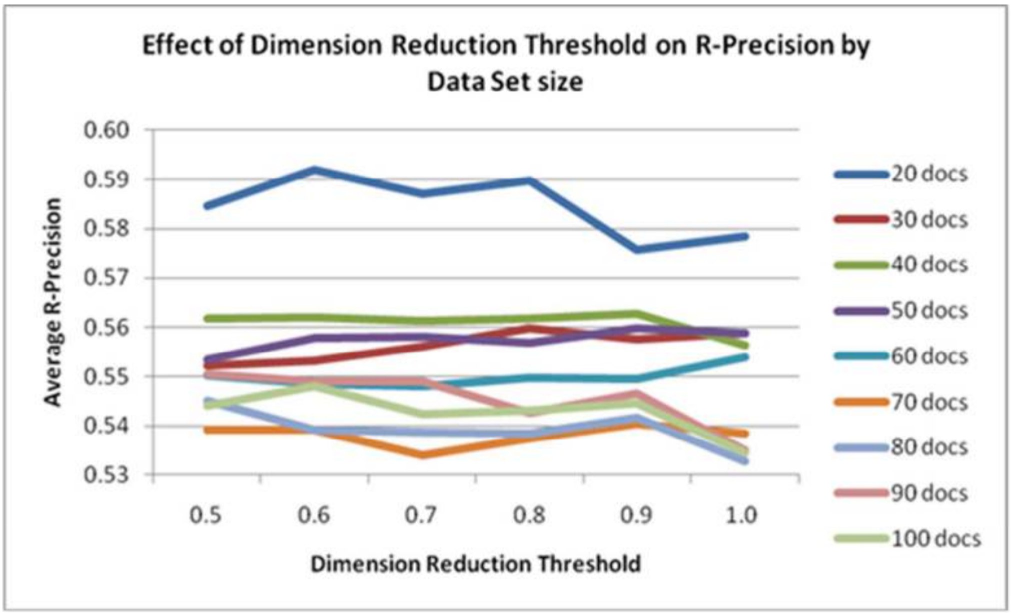


Figure 4. Effect of dimension reduction threshold on R-Precision

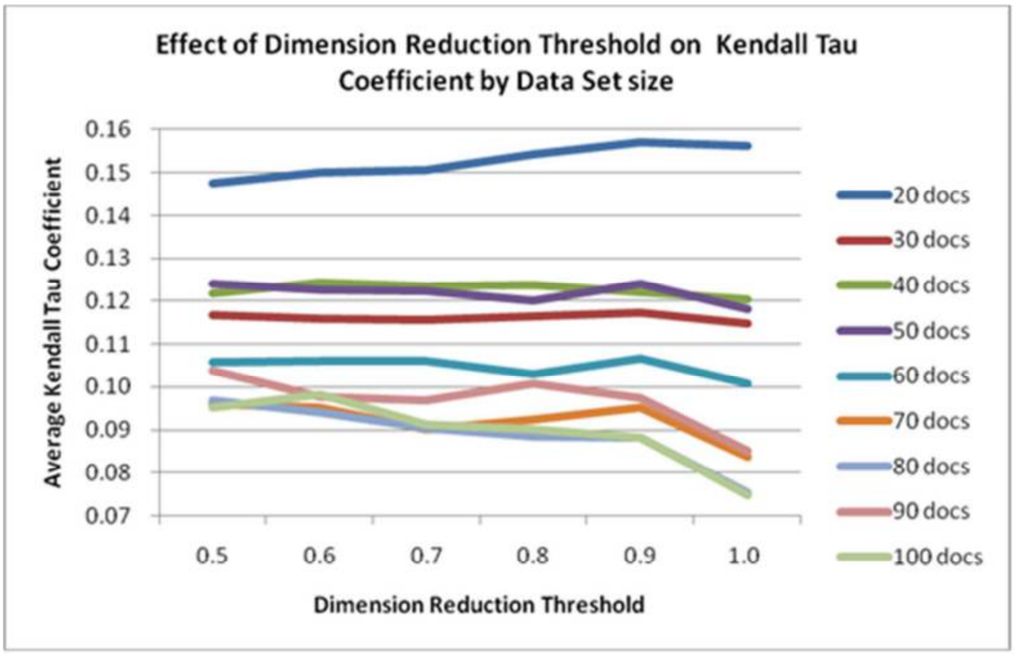


Figure 5. Effect of dimension reduction threshold on Kendall's tau coefficient

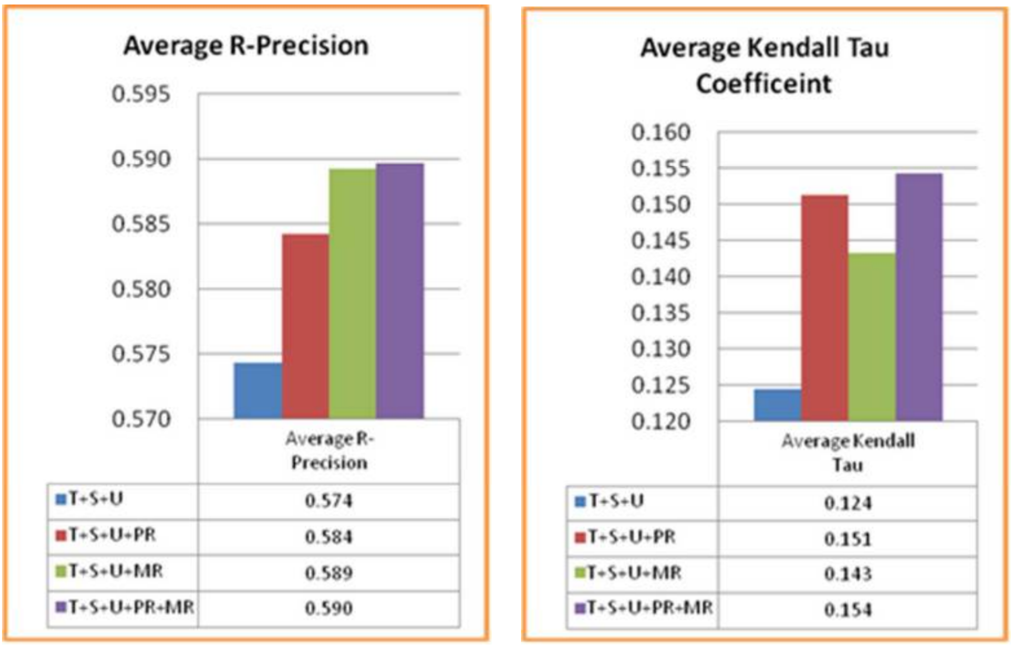


Figure 6. (a) Average R-Precision (b) Average Kendall's tau coefficient under the optimal parameters (dataset = 20 documents, dimension reduction threshold = 0.8)

4.4 Discussion

According to the experimental results, the maximum Kendall's tau value is only 0.154. Such a low value indicates that most of the new rank orders obtained with the proposed method are marginally similar with their original Google rank orders. Under such rank order disagreement, the proposed method achieved an R-precision close to 0.6, that is, about six of top ten documents remain within top ten after re-ranking. One of the typical examples for top 20 search results is shown below:

> The original rank order
> R1 = [d1, d2, d3, d4, d5, d6, d7, d8, d9, d10, d11, d12, d13, …, d20]
>
> The re-ranked order within top 10 positions
> R2 = [**d4, d6, d1,** *d14, d12***, d8**, *d11,* **d7, d3***, d13*, ….]

This case has an R-Precision of 0.6 because six documents d4, d6, d1, d8, d7, d3 are still ranked within top 10 positions, but the whole rank order is dissimilar to the original one, so it has a low Kendall tau coefficient of 0.155.

Obviously, the proposed method is still weak in estimating the Google rankings with a simple combination of only a few ranking factors. We intent to use more ranking factors and conduct weight adjustments in future research to reach better re-ranked orders such as the following cases:

R2 = [**d1, d3, d4, d6, d7, d8**, *d14, d13, d12, d11*,…], Kendall tau = 0.73, R-Precision = 0.6

R2 = [**d1, d3, d4, d6, d7, d8**, *d11, d12, d13, d14*, ….], Kendall tau = 1, R-Precision = 0.6

Moreover, we found the weights of factors achieved by the generic algorithm follow almost the same pattern. One set of the weights of factors that performed best in Kendall Tau and R-Precision under the proposed methods is: {W_T = 0.048, W_S = 0.01, W_U = 0.325, W_R = 0.617}. The results suggest that for the queries in the field of interest PageRank is the most important factor in estimating Google ranking. The document's URL follows as the second most important factor. Finally a document's title and its snippet have minor impact on search ranking. These findings cannot lead to generalization for queries from different fields as suggested in literature and in ranking factor surveys (Killoran , 2013; Mavridis & Symeonidis 2015; SEOmoz 2011, 2013).

For SEO practitioners interested in the industrial products field, we suggest they should pay special attention to PageRank to get better ranking. Additionally, the proposed method could be used to track current changes of the ranking algorithm Google applies. For example, we can set the genetic termination condition as R-Precision = 0.6 and then find the final weights of ranking factors of interest. Any change of weights can imply its change of impact on ranking. One of our recent simple test results indicate that the weight of URL factor is decreasing. This preliminary finding is consistent with our observation and that of recent ranking factor survey (Searchmetrics, 2015).

5. Conclusions

This study proposed an estimation function to approximate the ranking function of Google from an SEO perspective. The estimation function calculates a new query-document match score for each document using the weighted sum of query match scores from the title, snippet, and URL as well as off-page factors including PageRank and MozRank. The query match score counts the relevance scores of a set of terms semantically related a given query obtained via latent semantic analysis of the documents retrieved for each query. Additionally, the weights of each ranking factor were determined using a genetic algorithm. Experimental results indicate that the proposed method performed best in Kendall's tau coefficient and R-Precision when using a dataset of twenty documents with a dimension reduction threshold of 0.8. The results also reveal that the proposed method can help assess the probability of

SEO success with an R-precision of 0.6 while the ranks of the new and original result sets are quite dissimilar.

This study considered a limited set of ranking factors only. Although the proposed method produced satisfactory results in R-precision, there is still room for improvement in Kendall's tau coefficient. That is, to better approximate the original Google's ranking order with some more easily accessible ranking factors.

References

Abdi, H. (2007), *The Kendall Rank Correlation Coefficient*, In N. Salkind (Ed.), *Encyclopedia of Measurement and Statistics*, Thousand Oaks, CA: Sage Publications.

Agichtein, E., E. Brill, & Dumais, S. (2006), "Improving web search ranking by incorporating user behavior information", In *Proceedings of the 29th ACM SIGIR conference on Research and development in information retrieval*, pp. 19-26.

Al-Shboul, B, Myaeng, S.H. (2013), "Wikipedia-based query phrase expansion in patent class search", *Information Retrieval*, November 2013, pp. 1-22.

Baeza-Yates R. and Ribeiro-Neto, B. (2011), *Modern Information Retrieval: The Concept and Technology behind Search*, 2nd ed., New York: Addison-Wesley Professional.

Berman, R., and Katona, Z. (2013). The role of search engine optimization in search marketing. Marketing Science, 32(4), 644-651.

Bifet, A., C. Castillo, P. Chirita, & Weber, I. (2005), "An Analysis of Factors Used in Search Engine Ranking", In *Proceedings of First International Workshop on Adversarial Information Retrieval on the Web*, pp. 1-10.

Carpineto, C. and G. Romano. (2012), "A Survey of Automatic Query Expansion in Information Retrieval", *ACM Computing Survey*, Vol. 44, No.1, pp 1-50.

Chuklin, A. , Serdyukov, P. , De Rijke, M. (2013), "Modeling clicks beyond the first result page", In *Proceedings of International Conference on Information and Knowledge Management,* pp. 1217-1220.

Cutts, M. (2011, Aug. 17). Can you explain what Google means by "Trust"? Available at http://www.youtube.com/watch?v=ALzSUeekQ2Q, Google Webmasters Channel.

Derhami, V., Khodadadian, E., Ghasemzadeh, M., and Bidoki, A. M. Z. (2013), "Applying reinforcement learning for web pages ranking algorithms", *Applied Software Computing*, Vol. 13, No., 4, pp.1686–1692.

Enge, E., Spencer, S., Stricchiola, J. and Fishkin, R. (2012), *The Art of SEO: Mastering Search Engine Optimization*, 2nd ed., O'Reilly Media, Sebastopol, CA.

Evans, M. P. (2007), "Analysing Google rankings through search engine optimization data," *Internet Research*, Vol.17 No.1, pp.21-37.

Fortunato, S., Boguñá, M., Flammini A. and Menczer, F. (2008), "Approximating PageRank from In Degree", *Algorithms and Models for the Web Graph, Lecture Notes in Computer Science*, 4936, pp. 59-71.

Gandour, A. and Regolini, A. (2011), "Web site search engine optimization: a case study of Fragfornet", *Library Hi Tech News*, Vol. 28, No. 6, pp.6 – 13.

Gennaro, S. (2015), Brevity and clarity: Titles, key words, and search engine optimization, Journal of Nursing Scholarship, 47 (3), pp. 195-196.

Google Inc. (2010). *Google Search Engine Optimization Starter Guide*, available at http://www.google.com/webmasters/docs/Fsearch-engine-optimization-starter-guide.pdf

Hariri, N. (2011), "Relevance ranking on Google: Are top ranked results really considered more relevant by the users?", *Online Information Review*, Vol. 35, No. 4, pp.598 – 610.

Hopkins, L. (2012), Online reputation management: Why the first page of Google matters so much. available at http://www.leehopkins.net/2012/08/30/online-reputation-management-why-the-first-page-of-google -matters-so-much/. (accessed 26, May, 2014)

Kazembe, Chitsanzo W. (2011), Analyzing Google Search Results through Latent Semantic Analysis, Master's Thesis, Yuan Ze University, Taiwan.

Killoran, J.B. (2013), "How to Use Search Engine Optimization Techniques to Increase Website Visibility," *IEEE Transactions on Professional Communication,* vol. 56, no. 1, pp. 50-66.

Langville, L. N. and Meyer, C.D. (2006), *Google's PageRank and Beyond: the Science of Search Engine Rankings*, Princeton University Press, New Jersey and Oxford.

Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L. and Gay, G. (2006), "The influence of task and gender on search and evaluation behavior using Google", *Information Processing and Management,* Vol. 42 No. 4, pp. 1123-1131.

Manning, C., Raghavan F. and Schütze, H. (2009), *An Introduction to Information Retrieval*, Cambridge Universty Press, Cambridge, England.

Masri, A. M., Berrut, C., Chevallet, J. P. (2013), "Wikipedia-based semantic query enrichment", In *Proceedings of International Conference on Information and Knowledge Management,*, pp. 5-7.

Minhas, G. and Kumar, M. (2013), "LSI based relevance computation for topical web crawler", *Journal of Emerging Technologies in Web Intelligence*, Vol. 5, No. 4, pp. 401-406.

Mirzal, A. (2013), "The limitation of the SVD for latent semantic indexing", In *Proceedings of 2013 IEEE International Conference on Control System, Computing and Engineering*, pp. 413-416.

Moreno, L. and Martinez, P. (2013). "Overlapping factors in search engine optimization and web accessibility," *Online Information Review*, Vol. 37, No. 4, pp.564 – 580.

Patterson, A.L. (2009). Phrase-based indexing in an information retrieval system. *U.S. Patent 7,536,408 B2*, Washington, DC: U.S. Patent and Trademark Office.

Richardson, M., Prakash, A., and Brill, E. (2006), "Beyond PageRank: machine learning for static ranking", in *Proceedings of the 15th international conference on World Wide Web*, Edinburgh, Scotland, May 23–26, 2006, ACM, pp.707-715.

Sagot, S., Fougeres, A.-J., Ostrosi, E., Lacom, P., (2014), "Search engine optimization: From analysis based on an engineering meta-model towards integrative approaches," 2014 International Conference on Information Society (i-Society), pp.274-281.

Searchmetrics (2015), Search Ranking Factors 2015: Understand how the deck is stacked, available at http://www.searchmetrics.com/knowledge-base/ranking-factors/ (accessed August 6, 2015)

SEOmoz (2011), 2011 Search engine ranking factors, available at https://moz.com/search-ranking-factors/2011 (accessed August 6, 2015)

SEOmoz (2013). 2013 Search engine ranking factors, available at https://moz.com/search-ranking-factors/survey (accessed August 6, 2015)

Spink, A., Jansen, B.J, Blakely, C. and Koshman, S. (2006), "A study of results overlap and uniqueness among major Web search engines", *Information Processing and Management,* Vol. 42 No.5, pp. 1379-1391.

Su, A.J., Hu, Y. C., Kuzmanovic, A., and Koh, C.K. (2010), "How to improve your Google ranking: Myths and reality", In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'10).*

Torkestani, A. J. (2012), "An adaptive learning to rank algorithm: Learning automata approach", *Decision Support Systems*, Vol. 54, Issue 1, pp. 574-583.

Wall, M. E., Rechtsteiner, A. & Rocha, L. M. (2003), “Singular value decomposition and principal component analysis”, In D.P. Berrar, W. Dubitzky, & M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Kluwer, Norwell, MA, pp. 91-109.

Biographical Details

Cheng-Jye Luh is an Associate Professor in the Department of Information Management at Yuan-Ze University, Taiwan. His research interests include web mining and semantic web. Luh received a PhD in Electrical and Computer Engineering from University of Arizona, USA.

Sheng-An Yang received his Master’s degree in Information Management from Yuan-Ze University, Taiwan. His research interests include web mining and information retrieval.

Dean Ting-li Huang received his Master’s degree in Information Management from Yuan-Ze University, Taiwan. Huang received a BA degree in Information Management from Ming-Chuan University. His research interests are document clustering and semantic web applications.