Journal of Global Information Management
Volume 27 • Issue 1 • January-March 2019

A Study on Effective Measurement of Search Results from Search Engines

Jin Zhang, University of Wisconsin Milwaukee, Milwaukee, USA
Xin Cai, University of Wisconsin Milwaukee, Milwaukee, USA
Taowen Le, Weber State University, Ogden, USA
Wei Fei, Suzhou Library, Suzhou, China
Feicheng Ma, Wuhan University, Wuhan, China

ABSTRACT

This article describes how as internet technology continues to change and improve lives and societies worldwide, effective global information management becomes increasingly critical, and effective Internet information retrieval systems become more and more significant in providing Internet users worldwide with accurate and complete information. Search engine evaluation is an important research field as search engines directly determine the quality of information users' Internet searches. Relevance-decrease pattern/model plays an important role in search engine result evaluation. This research studies effective measurement of search results through investigating relevance-decrease patterns of search results from two popular search engines: Google and Bing. The findings can be applied to relevance-evaluation of search results from other information retrieval systems such as OPAC, can help make search engine evaluations more accurate and sound, and can provide global information management personnel with valuable insights.

KEYWORDS

Global Information Management, Information Retrieval, Internet Search Engine, Internet Search, Search Result Analysis, Search Result Evaluation, Search Result Measurement, Search Result Ranking

1. INTRODUCTION

As more and more people worldwide depend on the Internet to fulfill their information needs (Khatwani & Srivastava, 2017), and as the impact of Internet on people and societies have become increasingly profound (Teo, 2007; Lane et al., 2017), researchers throughout the world have studied factors maximizing successes of information technology implementations or global information management (Roztocki & Weistroffer, 2011, Lee et al., 2014; Caprio et al., 2015; Hung et al., 2016; Silic & Back, 2016; Soja, 2016; Chatterjee et al., 2017). One such technological implementation is the employment of search engines. Because of the critical role search engines play in bridging Internet information resources and information users, it is particularly important to evaluate effectiveness of search engines through effective measurements of their search results, as different search engines utilize different retrieval and ranking algorithms and therefore respond to search queries with different search results.

Average Internet searchers tend to take the search results presented by the search engines as a list of decreasing relevance, and they tend to browse only the first 20-30 items on a results list from a search engine. Moreover, business intelligence systems also seem to base many of their decisions

DOI: 10.4018/JGIM.2019010110

on search results as returned by Internet search engines. If the most relevant results are not properly positioned on the result list, important information would be missed, and the decisions could be impaired. Therefore, precise relevance ranking of search result items as returned by search engines is extremely important.

However, because what resides on the Web is an ever-changing and extremely heterogeneous data collection (Jansen & Pooch, 2001), Web page ranking algorithms have become very complicated and dynamic (Dean 2016; Barysevich 2017). It is important to know that ranking algorithms of different search engines handle variables differently. Consequently, the degree of search result relevance varies from search engine to search engine. Ideally, if all returned items are ranked in terms of relevance to the search query, and the ranked data are captured in a two-dimensional chart where the *X*-axis represents the ranked items and the *Y*-axis represents the relevance score, then a decline curve appears. Understanding the downward curve is critical to evaluating the quality of search results because the downward curve serves as a yardstick in measuring relevance of search results of a search engine.

The primary purpose of this study is to explore effective measurement of search results from search engines through investigating relevance-decrease patterns of search results from two major search engines: Google and Bing. To accomplish the purpose, 4 domain categories were defined, and 24 search queries with 6 from each category were formulated and submitted to both Google and Bing. Retrieved results were then collected, and their relevance was judged by 32 subjects independently. A group of possible regression models were developed for regression analysis, and the performances of the regression models were tested. The best-fit regression model was identified through *ANOVA* analyses. The findings of this study help people better understand the relevance-decrease patterns of search results produced by search engines. The best-fit regression model identified in this study provides a way for people to evaluate search result relevance of search engines.

2. RELATED RESEARCH

2.1. Importance of Search Engines

The Web has become a primary source of information due to the continued development of information and communication technology. It was reported that there were over 2.9 billion Internet users all over the world in 2014 (Internet live stats, 2015http://www.internetlivestats.com/internet-users/). Because of the richness and diversity of Web information (Zhang & Fei, 2010), it has become more and more challenging to efficiently and effectively search and find the needed information on the Internet. Fortunately, we have search engines to tackle this problem. Search engines have become the primary means in retrieving Web information. It was reported that over 91% of Internet users used search engines to find Web information and 54% of them were loyal users (who used search engines at least once a day) (Purcell, 2012). According to a more recent study (comScore, 2015), the top three search engines in use today are Google, Bing, and Yahoo, with respective market shares of 63.9%, 20.9%, and 12.5%, making Google and Bing the two most popular search engines in the world.

2.2. Relevant Aspects of Search Engines

The same search terms in a query constrained by different search features of a search engine such as returned result format or time period may result in different search results. Search engine interface design affects search results because the interface design has a strong impact on users' selection of search features. Retrieval algorithm of a search engine directly determines positions of returned items on a results list and is therefore the basis of search result ranking evaluation analysis. Search engine result ranking evaluations and methods require that a clearly-defined relevance measurement method be used in judging relevance of returned items on a results list.

2.2.1. Interface Design

Interface design is an important aspect of search engines. Most search engines have adopted natural language interface to facilitate user-friendly Internet searching, but it is more difficult for machines to understand natural-language queries than structured queries (Kaufmann & Bernstein, 2010). On the other hand, visualization technologies provide an intuitive interface, which can be used in modifying queries and discovering related topics (Tominski et al., 2009). Bilal (2002) emphasized that children are an important part of search engine users and criticized the design of Yahooligans for not considering enough children's needs.

2.2.2. Retrieval Algorithms

Retrieval algorithms are an essential component of search engines. Many retrieval algorithms have been developed with intent to reduce computational costs and storage expenses (Blei et al., 2003; Deerwester, et al., 1990; Hofmann, 1999).

2.2.3. Search Engine Evaluation Methods

Search engine evaluation has become an important research topic for many. The criteria of search engine evaluation can be multifold. While precision and recall have been widely used in the evaluation of traditional information retrieval systems, it is impossible to calculate the recall of a search engine because there is no way to find all the relevant items in the database of a search engine (Chu & Rosenthal, 1996). Precision, however, can still be used in the evaluation of search engines. Response time is another important measurement in evaluating search engines (Chignell & Gwizdka, 1999). Some introduced the criteria of search engine performance stability (Vaughan & Thelwall, 2004; Vaughan, 2004; Bar-Ilan, 1999, 2004). For example, researchers of Webometric collected data by using various search engines, and it is critical for them to know which engine is more stable than others (Thelwall, 2008). Hassan and Zhang (2001) compared image search engines using common features, but others believe that the availability of multiple languages is also an important criteria in search engine evaluation (Zhang & Lin, 2007; Davis, 1996; Gey et al., 2005).

Järvelinv and Kekäläinen (2002) proposed three search engines evaluation methods based on relevance scores of returned results. Direct Cumulated Gain adds the relevance scores of returned items directly to get an overall score from one query. Discounted Cumulated Gain (DCG) treated relevance scores differently by looking at the position of items. In other words, the relevance score would be reduced as the rank of the item increases. DCG is an important measurement in comparing search engines (Carterette & Jones, 2008; Zheng et al., 2007; AI_Maskari et al., 2007; Lin et al., 2013; Wu, 2011; da Costa Pereira et al., 2012; and Zhou & Yao, 2010). Defined as reciprocal of the logarithm of its rank (Järvelin & Kekäläinen, 2002), the discounted equation is usually used in describing the downward trend of search result relevance in these studies. The function reduces the importance of lower-ranked items because they have a smaller chance to be viewed than the higher-ranked ones. Normalized Discounted Cumulated Gain (nDCG) is a standardized variant of the DCG method, but no specific studies on nDCG were found.

2.2.4. Search Result Ranking Algorithm

Although there are many methods for search engine evaluation, result ranking remains one of the most important methods for two reasons. First, there is always a gap between the returned results and the users' information needs. No matter how accurate the applied ranking algorithms are, there are always differences between ranking by the search engine and ranking by the users even if the users have the same foundation knowledge of a topic (Bar-Ilan et al., 2007). The researchers contributed this discrepancy to the cognitive, affective, and physical factors. Xie (2000) ascribed this difference to the interactive and ever-changing nature of the information retrieval process. Secondly, most users only view one results page (Jacsó, 2008; Spink & Jansen, 2004), and it is proven that they tend to

view fewer documents over times (Jansen & Spink, 2006). Consequently, it is particularly important for search engines to satisfy user's information needs by showing the most relevant results on the top of the returned result list.

Thus, employment of sound result-ranking algorithms is critical to search engines. PageRank is the most famous ranking algorithm which borrows the idea from citation analysis (Brin & Page, 2012). Usually relevance would be computed by a similarity method such as cosine similarity method or distance similarity method (Zhang & Korfhage, 1999; Wilbur & Sirotkin, 1992). Ranking by recentness ensures that users receive the most recent information (Efron & Golovchinsky, 2011). Zhang and Dimitroff (2004) noted that Webpage content characteristics, such as keyword position, layout, and keyword frequency, can affect the ranking of a Webpage.

2.3. Summary

In summary, there have been many studies on search engine evaluations, search engine feature analysis, and search engine results ranking evaluation, but studies on relevance-decrease patterns of search results, which are critical to search engine results ranking are scant in the literature.

3. RESEARCH METHODOLOGY

There are usually multiple items on the results list returned by a search engine in response to a query. The items on the list are retrieved from the database because they are determined by the search algorithm to be relevant to the query submitted to the search engine. The relevance of one item on the list is different from the relevance of another item to some degree if these two items are not identical (usually duplicated items are excluded from a returned results list). The breadth, depth, style, and emphasis of the two items may differ although both might address the same topic. If relevance of each item on the list is evaluated and assigned a relevance score, all the items on the list can be ranked in descending order of relevance scores as done by most search engines. The items on the ranked results list would follow a downward relevance trend or pattern in terms of relevance. Such pattern can be revealed if proper regression analyses are applied to ranked results sets.

To investigate relevance-decrease patterns of search results, data were collected from the two best-known search engines, namely Google and Bing. To ensure data representativeness, search tasks or queries were designed from 4 major subject categories: Health, News and Media, Science and Technology, and Economy and Business.

As part of the data collection process, search queries covering various subject categories were formulated and submitted to search engines. The result set from each query as returned by each search engine was captured and saved. To minimize subjectivity in human judgements, the result sets were presented to a total of 32 evaluators. All evaluators were college students as modern college students were typically familiar with Internet search systems. 16 of them were students of Suzhou University in China, including 6 males and 10 females; the other 16 were students of Weber State University in the U.S., including 10 males and 6 females. They were randomly approached but must meet three basic requirements to be selected: (1) must be willing to participate (to ensure serious analysis), (2) must be at least a junior (to ensure a decent knowledge base), and (3) must be proficient in English (as search terms and retrieved pages were in English). Since they were randomly approached, their majors covered a large variety of fields.

Each evaluator independently evaluated the result sets and scored each item on the result sets in terms of relevance to the query statement. The relevance scores as assigned by the 32 evaluators were then plugged into different regression models to identify the best-fit model at a reasonable significance level. The best-fit regression model was then used to describe the relevance-decrease patterns of search results.

To provide direction for the study, the following 5 null hypotheses were proposed:

($H1_0$): There are no significant differences among the relevance regression equations or models in terms of R^2 in the search results.
($H2_0$): There are no significant differences among the 4 subject categories in terms of R^2 in the search results.
($H3_0$): There are no significant interactions between the relevance regression models and the subject categories in terms of R^2 in the search results.
($H4_0$): There are no significant differences between the two search engines in terms of R^2 in the search results.
($H5_0$): There are no significant interactions between the relevance regression models and the search engines in terms of R^2 in the search results.

In $H1_0$ the independent variable is relevance regression models, and the dependent variable is R^2. It is the primary hypothesis for this study. The result of this test is used to identify the best-fit model in describing the relevance-decrease trend of search results from a search engine.

In $H2_0$ the independent variable is subject field, and the dependent variable is R^2. This hypothesis examines whether the models vary in different subject categories. In other words, it examines whether the nature of subject categories affects the selection of a best-fit regression model.

In $H3_0$ the independent variables are subject field and relevant regression equation, and the dependent variable is still R^2. This hypothesis examines whether there are interactions between the two independent variables.

In $H4_0$ the independent variable is search engine, and the dependent variable is R^2. It is necessary to investigate whether there is any difference between the two search engines in terms of regression analysis results.

In $H5_0$ the independent variables are subject field and search engine, and the dependent variable is still R^2. This hypothesis examines whether there are interactions between the two specified independent variables.

The significance level (α) for the testing of all these hypotheses was set to 0.05. In other words, if a produced *p*-value from an inferential test is larger than 0.05, the corresponding hypothesis is accepted; otherwise, the hypothesis is rejected.

The detailed data collection, relevance-score analysis, and regression analysis methods are discussed respectively as follow.

3.1. Data Collection

As mentioned earlier, Google (2016) and Bing (2016) were widely regarded as the most popular Internet search engines (comScore, 2015). This study employed both of them.

Search queries covered 4 subject categories: *Health, News and Media, Science and Technology,* and *Economy and Business*. These categories are similar to some of the categories in Yahoo Directory but with some revision (Yahoo Directory, 2016). These 4 categories were selected to cover diverse domain areas to reflect different natures of search tasks. It suggests that subject categories selected represent information needs of common people. In each subject field, 6 search tasks or queries covering were carefully designed to represent each field. For instance, popular topics such as *Lady Gaga*, *Obamacare, Bin Laden death, The Korean crisis*, *The Syria crisis*, and *H7N9 bird flu* were included in the field of *News and Media*. Specific search tasks in each subject field are listed in Table 1 where strings in parentheses represent IDs for subject categories or for search tasks (for instance, *C1_T1* represents Autism in Health).

Previous studies (Zhang & Fei, 2010; Zhang, Fei & Le, 2013) suggest that the ranking pattern of retrieved items on a returned results list becomes stabilized when the size of the returned results list reaches 50. In addition, users are only interested in the top 20 items from a returned results list (Jacsó, 2008; Jansen & Spink, 2006). Therefore, for each search task in this study, only the first 50 records from each search result list were selected for analysis. Titles and Webpage contents of the

Table 1. Summary of the Subject Categories and Search Tasks

Health (C1) Queries 1-6	News and Media (C2) Queries 7-12	Science and Technology (C3) Queries 13-18	Economy and Business (C4) Queries 19-24
Autism (C1_T1) Q1	Lady Gaga (C2_T1) Q7	Google glasses (C3_T1) Q13	BRICS (C4_T1) Q19
Weight control (C1_T2) Q2	Obamacare (C2_T2) Q8	Global warming and climate change (C3_T2) Q14	World Trade Organization (C4_T2) Q20
Smoking and health (C1_T3) Q3	Bin Laden death (C2_T3) Q9	Web 2.0 (C3_T3) Q15	US dollar and Chinese Yuan exchange rate (C4_T3) Q21
AIDS prevention (C1_T4) Q4	The Korean crisis (C2_T4) Q10	Wind energy (C3_T4) Q16	Hedge Fund (C4_T4) Q22
Asthma (C1_T5) Q5	The Syria crisis (C2_T5) Q11	Electric car (C3_T5) Q17	The Big Mac Index (C4_T5) Q23
Birth control (C1_T6) Q6	H7N9 bird flu (C2_T6) Q12	Stem cell research (C3_T6) Q18	Micro-economy (C4_T6) Q24

Each search task corresponded to a search query. Each query was submitted to both Google and Bing.

50 records for each search task were recorded for later relevance analysis. Ranking information of retrieved items was not relevant to this study; therefore, it was not presented to the evaluators.

With 6 search tasks for each subject category, a total of 24 (4×6=24) search result or retrieval data sets (DS) were created. Since each search task was submitted to both Google and Bing, each retrieval data set contained two subsets: one for Google, and the other for Bing.

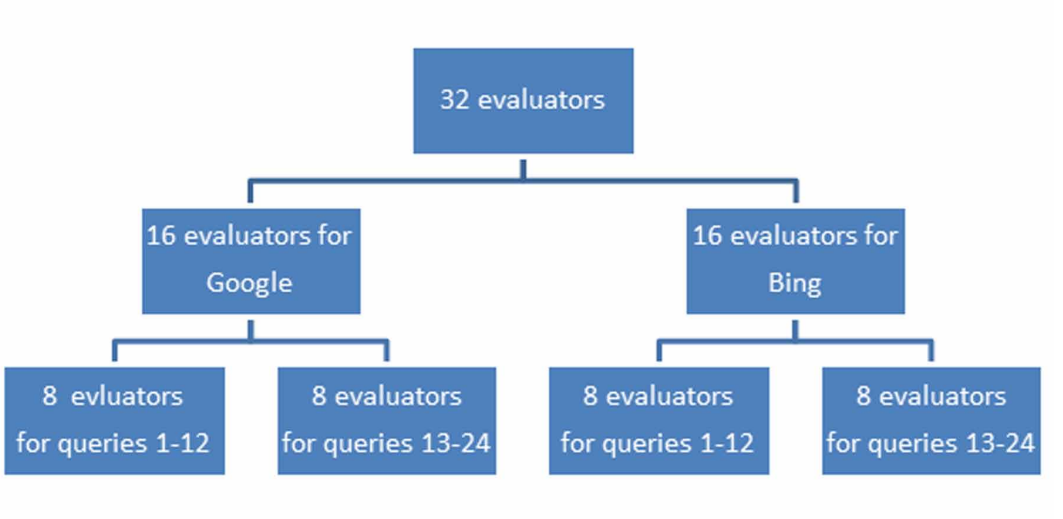
3.2. Relevance Judgment

The 24 data sets collected were then presented to 32 individual evaluators for relevance judgment. Each data set was randomly assigned to 8 evaluators for independent evaluation. The researchers provided the evaluators with score assignment instructions which required all evaluators to read the content of the provided records (that is the retrieved Webpages saved at the time of query submission) before assigning their scores. The researchers also monitored the score assigning process. If any evaluator had questions on the assigned search tasks, the researchers would answer the questions and provide any needed clarification. Based on the perceived relevance of the record to the search task, the evaluator would assign a relevance score to each item.

The score system was based on an 11-point scale with 0 for totally irrelevant, 2 for largely irrelevant, 3 for basically irrelevant, 4 for somewhat irrelevant, 5 for somewhat relevant, 6 for basically relevant, 7 for largely relevant, 8 for relevant, 9 for very relevant, 10 for most relevant. Evaluators can assign to each item a score between 0 and 10.

The 32 evaluators were divided into two groups, one group (16 evaluators) for the Google data subsets, and the other (also 16 evaluators) for the Bing data subsets. In each group, 8 of 16 evaluators evaluated the first 12 of the 24 retrieval data sets while the other 8 evaluators evaluated the remaining 12 retrieval data sets. The evaluators were randomly assigned to the retrieval data sets. The relationships among the search engines, search tasks, and evaluators are shown in Figure 1.

Figure 1. Task Assignment



Since each record in a retrieval data set was evaluated by 8 different evaluators, it received 8 relevance scores. The average of the 8 relevance scores was calculated and used as the final relevance score for that record.

With 24 search queries formulated and submitted to each of the 2 search engines, and with each retrieval data set containing 50 records and evaluated by 8 evaluators, a total of 384 (24×2×50×8=19200) relevance scores were produced in this study.

3.3. Generation of Regression Models

As mentioned earlier, the first 50 records returned by a particular search engine in response to each search task were captured in the data set for that search task. To minimize subjectivity impact, each record was evaluated by 8 individual evaluators, and the average relevance scores of 8 evaluators for each record was calculated. Then the average scores were ranked in descending order. These data were used as input raw data in later regression analysis.

The regression analysis was based on two principles: First, if the curve characteristics of a regression equation best match the characteristics of the downward relevance trend of the items on the retrieved data sets, the regression model would be selected. Secondly, given relevant requirements met, only the simple and straightforward regression models would be chosen.

Following these principles, a group of 8 potential regression models were proposed for this study. After the regression analysis for each of the selected regression models, the best-fit regression model was identified. The following were the 8 equations or models proposed:

$$RM1(X) = b0 + \frac{b1}{X} \tag{1}$$

In Equation (1), *b0* and *b1* are constants which are larger than 0; $X \geq 1$.

$$RM2(X) = b0 + \frac{b1}{X^{1/2}} \tag{2}$$

In Equation (2), *b0* and *b1* are constants which are larger than 0; $X \geq 1$.

$$RM3(X) = b0 + \frac{b1}{X^{1/4}} \tag{3}$$

In Equation (3), *b0* and *b1* are constants which are larger than 0; $X \geq 1$.

$$RM4(X) = b0 + \frac{b1}{X^{1/8}} \tag{4}$$

In Equation (4), *b0* and *b1* are constants which are larger than 0; $X \geq 1$.

$$RM5(X) = b0 + \frac{b1}{Log_{2}(X)} \tag{5}$$

In Equation (5), *b0* and *b1* are constants which are larger than 0; $X \geq 1.5$.

$$RM6(X) = b0 + \frac{b1}{Ln(X)} \tag{6}$$

Equation (6) is similar to Equation (5) except the base of the logarithm function. Here $X \geq 1.5$.

$$RM7(X) = b0 + \frac{b1}{Log_{10}(X)} \tag{7}$$

In Equation (7), *b0* and *b1* are constants which are larger than 0; $X \geq 1.5$.

$$RM8(X) = b0 + \frac{b1}{2^{X}} \tag{8}$$

In Equation (8), *b0* and *b1* are constants which are larger than 0; $X \geq 1$.

All these models corresponded to a downward-trend curve which met the requirements for the regression model selection.

3.4. Hypothesis Testing

The processed data from the relevance judgment of each search task or query were plugged into the 8 regression equations respectively, and results such as corresponding parameters (*b0* and *b1*) and R^2 were collected from the regression analysis. Here R^2 is defined as 1 minus the ratio of residual sum of squares to corrected sum of squares. It indicates how well a resultant regression curve matches an input data set. Its valid value falls between 0 and 1. The larger the R^2, the better the corresponding regression model fits the data set; and vice versa. For each search engine, the 8 regression analyses were conducted for each of the 24 search tasks in the defined categories to obtain resultant data. Since each task was executed on two different search engines and each regression model corresponded to 24

individual tasks, each regression model produced 48(2×24=48) resultant R^2. Since the 8 regression models were being examined, there were a total of 384 (48×8=384) resultant R^2.

A two-factor *ANOVA* test was conducted to test the proposed null hypotheses $H1_0$, $H2_0$, and $H3_0$. The significance level for the test was set to 0.05. If a produced *p*-value from an inferential test result was larger than 0.05, the corresponding hypothesis would be accepted. Otherwise, the hypothesis would be rejected. If it was rejected, then a follow-up *Tukey* test would be used to detect which regression models and/or subject categories caused the rejection. The *ANOVA* test in conjunction with a follow-up *Tuke*y test would identify the best-fit regression model(s) if there were significant differences among the regression equations/models and the subject categories in terms of R^2. In the *ANOVA* test, the independent variables were regression equation/model and subject category while the dependent variable was R^2.

Another two-factor *ANOVA* test was conducted to test the proposed null hypotheses $H4_0$ and $H5_0$. The significance level for the test was also set to 0.05. The *ANOVA* test in conjunction with a follow-up *Tuke*y test would identify the reasons of hypothesis rejection if there were significant differences among the regression equations/models and the search engines in terms of R^2. In the *ANOVA* test, the independent variables were regression equation/model and search engine while the dependent variable was R^2.

The statistics software package *SPSS* (Version 20) was used for the regression analyses and the two-factor *ANOVA* analyses.

4. RESULTS AND DISCUSSIONS

4.1. The Descriptive Summary

4.1.1. The Descriptive Summary of the Relevance Scores

The raw data were collected from the 32 evaluators who evaluated the relevance of various search result sets. As stated earlier, a total of 19200 relevance scores were collected in this study.

Table 2 shows the descriptive summary of all the relevance scores. Each subject category received 4800 individual scores. The mean score for *Science and Technology* (C3) is the largest (7.0190), which indicates that the retrieved results from C3 are the most relevant compared to search results of the other categories. The means of the other three categories are, in a descending order, *Health* (C1) (6.9681), *News and Media* (C2) (6.9108), and *Economy and Business* (C4) (6.8512). The largest standard deviation pertains to *Economy and Business* (C4), 2.24911. The standard deviation for *Science and Technology* (C3) is the lowest (2.15019).

The mean relevance scores for Google is 6.8373, which is smaller than that of Bing (7.0373). The standard deviation for Google (2.28799) is higher than that for Bing (2.09700). The mean of the total relevance scores is 6.9373 and the standard deviation is 2.19679. Figure 2 shows the mean relevance scores in the four subject categories; the *Y*-axis represents mean relevance scores, and the *X*-axis represents the different search engines. It shows that Google performed better in *Science and Technology* and *Economy and Business* while Bing outperformed Google in *Health* and *News and Media*.

Table 3 exhibits the distribution of the relevance score frequencies for the entire data collection. In Table 3, the columns represent the frequency of each relevance score. The most frequently-occurring score is 8 (3802 times), which accounts for 19.80% of all the evaluations. The least-occurring score is 0, which only appears 7 times.

Figure 3 was produced to show relevance-decreasing patterns in a more intuitive way. It shows the descending curves of the relevance scores in the 4 subject categories. There are 4 subfigures in Figure 3 and each subfigure represents a subject category. In each figure, the *X*-axis captures the relevance ranks of the returned items while the *Y*-axis shows the relevance scores received. In this study, each subject category has 6 search tasks or queries, and each search task or query was represented by a

Table 2. The Descriptive Summary of the Relevance Scores

Category	Search Engine	Mean	Min	Max	Standard Deviation	Number of Relevance Scores
Health (C1)	Google	6.5913	0	10	2.35262	2400
	Bing	7.3450	0	10	1.94916	2400
	Total	6.9681	0	10	2.19274	4800
News and Media (C2)	Google	6.6375	0	10	2.29782	2400
	Bing	7.1842	1	10	2.04316	2400
	Total	6.9108	0	10	2.19111	4800
Science and Technology (C3)	Google	7.1000	1	10	2.23299	2400
	Bing	6.9379	1	10	2.06135	2400
	Total	7.0190	1	10	2.15019	4800
Economy and Business (C4)	Google	7.0204	1	10	2.22307	2400
	Bing	6.6821	1	10	2.26270	2400
	Total	6.8512	1	10	2.24911	4800
Total	Google	6.8373	0	10	2.28799	9600
	Bing	7.0373	0	10	2.09700	9600
	Total	6.9373	0	10	2.19679	19200

Figure 2. Mean Relevance Scores for the 4 Subject Categories

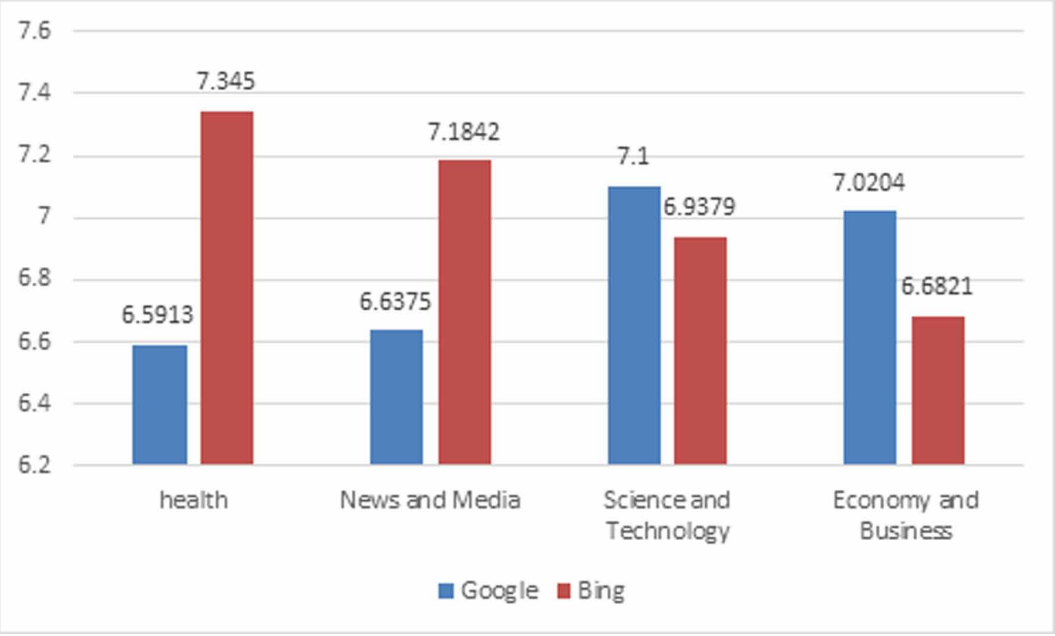


Table 3. The Distribution of Relevance Score Frequencies

Frequency	0	1	2	3	4	5	6	7	8	9	10	Total
Number	7	502	422	483	975	2268	2720	3067	3802	2584	2370	19200
Percent	0.04	2.61	2.20	2.52	5.08	11.81	14.17	15.97	19.80	13.46	12.34	100.0

curve. For example, *C1_T1* in *Health* (C1) represents the descending curve of search task or Query 1 in the subject category of *Health*.

4.1.2. The Descriptive Summary of the Results for the Regression Models

Since 24 search tasks were executed on each of the two search engines, and relevance scores for each search task on each engine was plugged into each of the 8 proposed regression models, a total of 384 (24×2×8) R^2 were generated. In Table 4, Regression Models 1 to 8 represent Regression Equations (1) to (8), respectively. In terms of mean R^2 values, Equation (4) (0.7899), Equation (3) (0.7477), Equation (2) (0.6322), Equation (1) (0.4698), Equation (5) (0.3723), Equation (6) (0.3723), Equation (7) (0.3723), and Equation (8) (0.2519) achieved the respective positions of first, second, third, fourth, fifth, sixth, seventh, and eighth.

Figure 4 shows the R^2 score means for the regression models. In Figure 4, the *X*-axis represents the regression models while the *Y*-axis represents the R^2 score means.

4.1.3. The Descriptive Summary of the Results from the 4 Subject Categories

Table 5 shows the descriptive summary of R^2 for the 4 subject categories. In Table 5, the rows represent the R^2 score means for the 4 subject categories while the columns represents the 8 regression models. Equation (4) (or Model 4) outperformed the other regression equations in terms of R^2 in *Health* (C1) (0.7745), *News and Media* (C2) (0.8151), *Science and Technology* (C3) (0.7689), and *Economy and*

Figure 3. Relevance Scores in the 4 Subject Categories

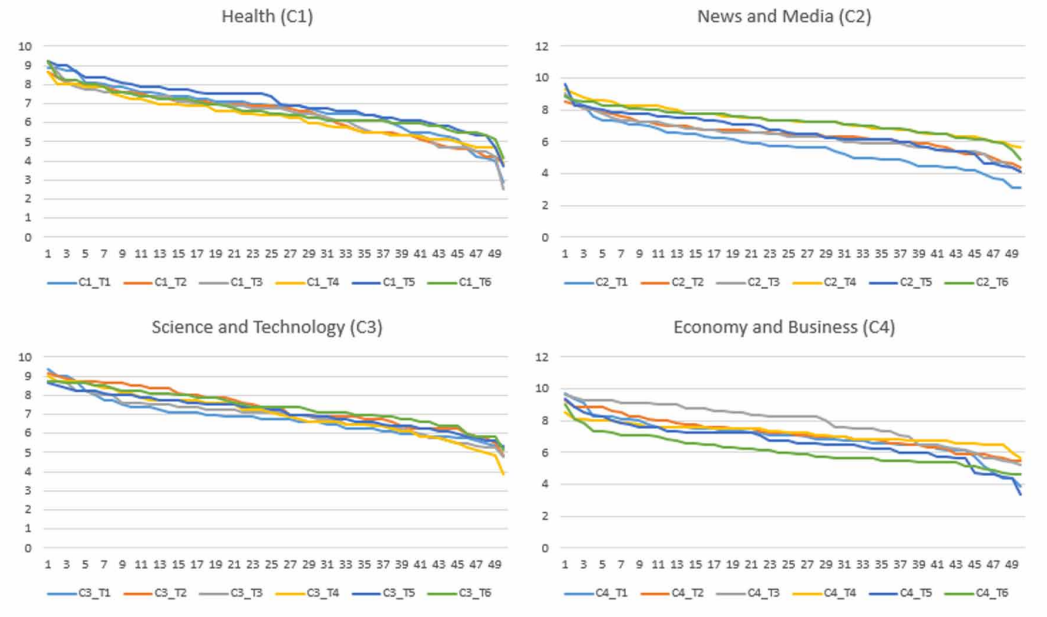


Table 4. The Descriptive Summary of R^2 for the 8 Regression Models

Regression Model	Mean	Std. D	N
1	0.4698	0.07566	48
2	0.6322	0.09618	48
3	0.7477	0.07091	48
4	0.7899	0.06694	48
5	0.3723	0.06728	48
6	0.3723	0.06728	48
7	0.3723	0.06728	48
8	0.2519	0.06016	48
Total	0.5010	0.19870	384

Business (C4) (0.8009). In other words, Equation (4) achieved the best performance in all subject categories.

Figure 5 displays the R^2 score means for the 4 subject categories. In Figure 5, the *X*-axis represents the regression models while the *Y*-axis represents the R^2 score mean. Each curve represents a corresponding subject category. It is clear that each curve reaches its peak at Equation (4).

4.1.4. The Descriptive Summary of the Results for the 2 Search Engines

Table 6 shows the descriptive summary of R^2 for the search engines. In Table 6, the columns represent the regression models while the row represents the search engines. Google achieved better performances across all regression models or equations except Equation (2).

Figure 4. R^2 Score Means for the 8 Regression Models

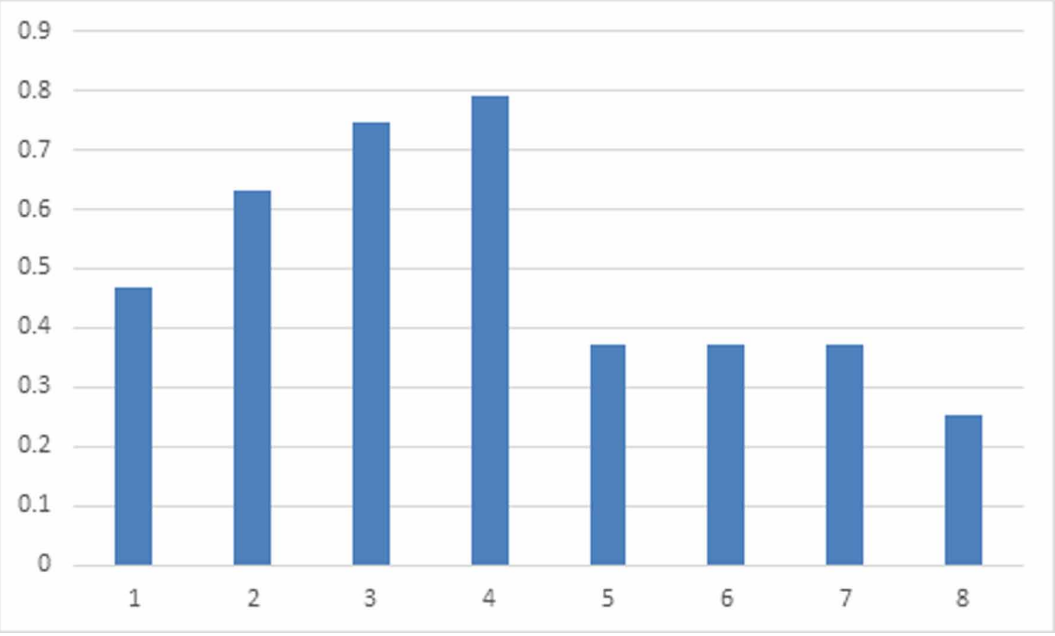


Table 5. The Descriptive Summary of R^2 for the Subject Categories

Domain Area	Equation (1)	Equation (2)	Equation (3)	Equation (4)	Equation (5)	Equation (6)	Equation (7)	Equation (8)	*Total*
C1	0.4614	0.6410	0.7328	0.7745	0.3682	0.3682	0.3682	0.2494	0.4955
C2	0.4959	0.6819	0.7742	0.8151	0.3953	0.3953	0.3953	0.2705	0.5279
C3	0.4380	0.6257	0.7240	0.7689	0.3408	0.3408	0.3408	0.2247	0.4755
C4	0.4837	0.5800	0.7599	0.8009	0.3849	0.3849	0.3849	0.2632	0.5053

Figure 5. R^2 Score Means for the 4 Subject Categories

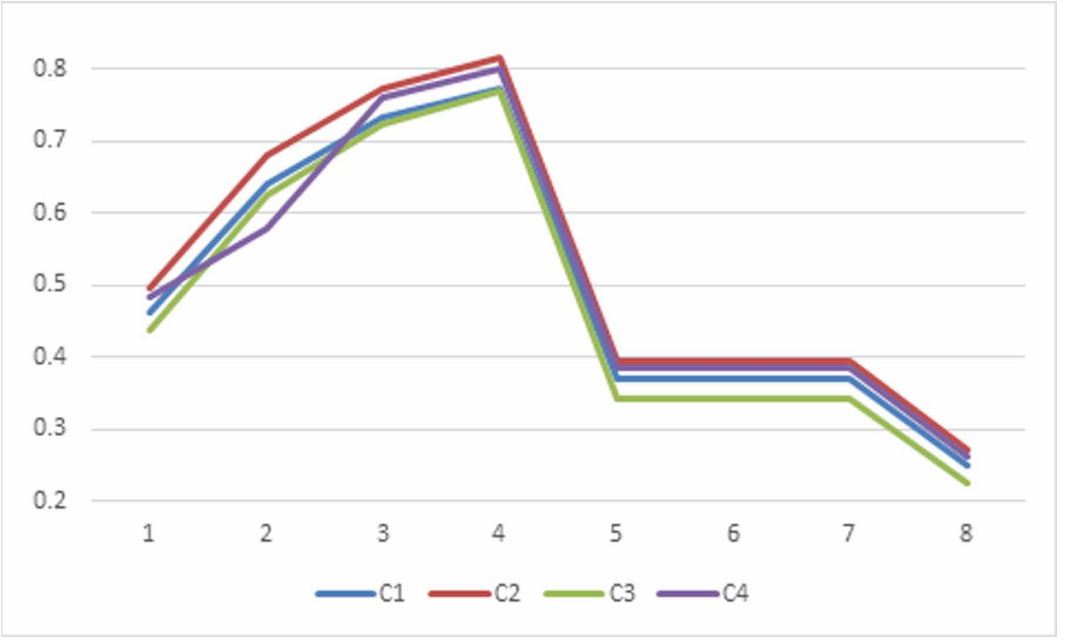


Table 6. The Descriptive Summary of R^2 for the Search Engines

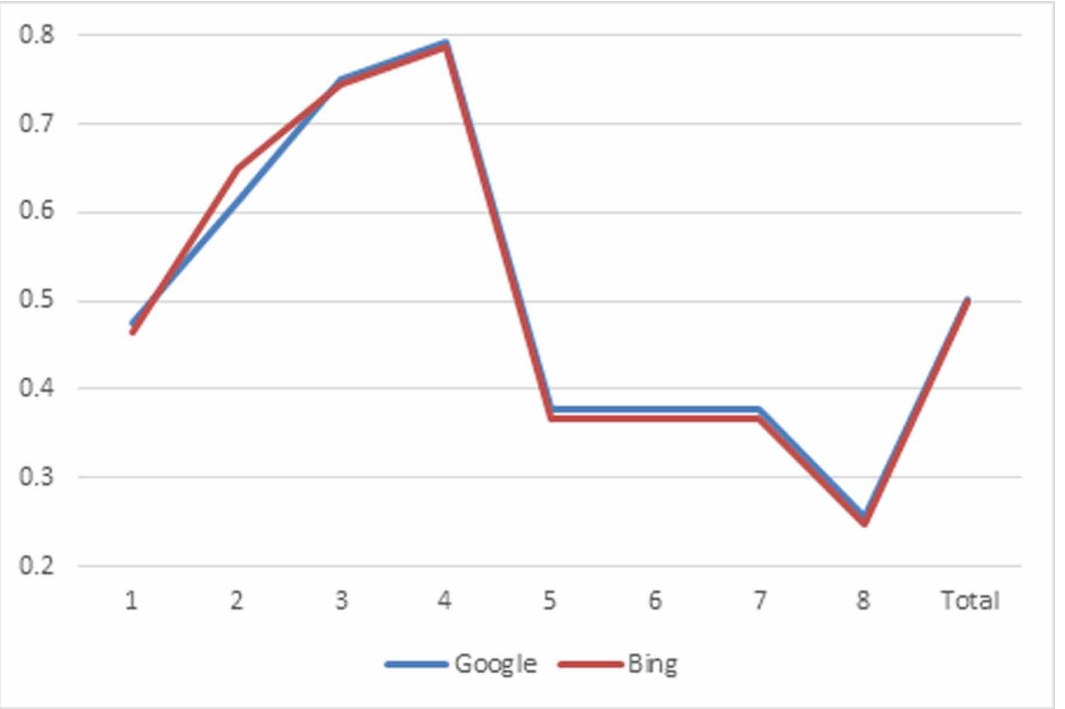
Search Engine	Equation (1)	Equation (2)	Equation (3)	Equation (4)	Equation (5)	Equation (6)	Equation (7)	Equation (8)	Total
Google	0.4743	0.6135	0.7498	0.7915	0.3771	0.3771	0.3771	0.2568	0.5021
Bing	0.4652	0.6508	0.7456	0.7883	0.3675	0.3675	0.3675	0.2471	0.5

Figure 6 displays the R^2 score means for the 2 search engines across the 8 regression models. In Figure 6, the *X*-axis represents the regression models while the *Y*-axis represents the R^2 score mean. Each curve represents a search engine.

4.2. Inferential Statistics Analysis

The hypotheses were proposed to examine the performances of the regression models, search engines, subject categories, and their interactions. Two two-factor *ANOVA* tests were conducted to test these hypotheses. The first was to test $H1_0$, $H2_0$ *and* $H3_0$; these three hypotheses were proposed to examine

Figure 6. R^2 Score Means for the Search Engines



the performances of the regression models, subject categories, and their interactions. The second was to test $H4_0$ and $H5_{0:}$ these two hypotheses were proposed to examine the performances of the regression models, search engines, and their interactions.

4.2.1. Results for Hypotheses $H1_0$, $H2_{0,}$ and $H3_0$

Since $H1_0$, $H2_{0,}$ *and* $H3_0$ were tested by the first two-factor *ANOVA test*, the test results for these three hypotheses are reported together. Table 7 shows the results. For hypothesis $H1_0$, with $df\left(7,352\right)$, the critical value at significant level (0.05) is 2.04, and *p*-value is 0.000. The critical value is much smaller than the *F* value of $H1_0$ (376.165) and *p*-value is also smaller than the significant level (0.05). It suggests that $H1_0$ is rejected and there are significant differences among the regression models in terms of R^2. For Hypothesis $H2_0$, with $df\left(3,352\right)$, the critical value at significant level (0.05) is 2.63, smaller than the *F* value of $H2_0$ (9.121), and *p*-value (0.000) is smaller than the significant level (0.05). Therefore, $H2_0$ is rejected, and there are significant differences among the 4 subject categories. For hypothesis $H3_0$, with $df\left(21,352\right)$, the critical value at significant level (0.05) is 1.56, which is bigger that the *F* value of $H3_0$ (0.599), and the *p*-value (0.919) is larger than the significant level (0.05). It suggests that $H3_0$ is accepted and there is no significant interaction between the regression equations and the subject categories in terms of R^2 .

Since $H1_0$ and $H2_0$ were rejected, two follow-up *Tukey* tests were conducted to detect the reasons of the respective rejections. Table 8 shows the results of the follow-up *Tukey* test for the regression models ($H1_0$). In this table, *I* and *J* stand for regression models. Mean difference (*I-J*) represents the difference between the R^2 scores of two regression models *I* and *J*. For example, mean difference (1-2) is the difference between the R^2 score means of Equations (1) and (2). Mean differences with an asterisk indicate significant differences. For example, the mean differences of Equation (1) and Equations (2)

Table 7. The Results for $H1_0$, $H2_0$, and $H3_0$

Factor	Type III Sum of Squares	df	F	Sig.
Regression Models	13.162	7	376.165	0.000
Domain categories	0.137	3	9.121	0.000
Interactions	0.063	21	0.599	0.919

to (8) are -0.1624*, -0.2780*, -0.3201*, 0.0974*, 0.0974*, 0.0974*, 0.2178*, respectively. Therefore, the differences between the following regression models {Equations (1) and (2) (-0.1624*), Equations (1) and (3) (-0.2780*), Equations (1) and (4) (-0.3201*), Equations (1) and (5) (0.0974*), Equations (1) and (6) (0.0974*), Equations (1) and (7) (0.0974*), Equations (1) and (8) (0.2178*), Equations (2) and (3) (-0.1156*), Equations (2) and (4) (-0.1577*), Equations (2) and (5) (0.2599*), Equations (2) and (6) (0.2599*), Equations (2) and (7) (0.2599*), Equations (2) and (8) (0.3802*), Equations (3) and (5) (0.3754*), Equations (3) and (6) (0.3754*), Equations (3) and (7) (0.3754*), Equations (3) and (8) (0.4958*), Equations (4) and (5) (0.4175*), Equations (4) and (6) (0.4175*), Equations (4) and (7) (.4175*), Equations (4) and (8) (.5379*), Equations (5) and (8) (.1204*), Equations (6) and (8) (0.1204*), and Equations (7) and (8) (0.1204*)} caused the rejection of $H1_0$.

Table 9 shows the results of the follow-up *Tukey* test for $H2_0$. In this table, *I* and *J* stand for the subject categories. Mean difference *(I-J)* represents the difference between the R^2 score means of two categories *I* and *J*. The differences between the following subject categories {*Health* and *News and Media* (-.0324*), *News and Media* and *Science and Technology* (.0524*), *Science and Technology* and *Economy and Business* (-.0298*)} led to the rejection of $H2_0$.

In summary, Equation (4) surpassed the other equations in each of the 4 subject categories. The hypothesis tests suggest that Equation (4) is the best-fit model among the 8 regression models proposed across all 4 subject categories.

4.2.2. Results for Hypotheses $H4_0$ and $H5_0$

Since the second two-factor *ANOVA* test was conducted to test $H4_0$ and $H5_0$, test results for both hypotheses were reported together. These hypotheses were to examine the performances of the regression models, search engines, and their interactions. Table 10 shows the results.

For $H4_0$, the critical value with $df\left(1,368\right)$ and at significance level of 0.05 is 3.87, which is much larger than the *F* value of $H4_0$ (0.086). Resultant *p*-value is 0.077 and is larger than the significant level (0.05). It indicates that $H4_0$ is accepted and there are no significant differences between the search engines in terms of R^2. As for interactions between the regression models and the search engines, with $df\left(7,368\right)$, the critical value at significant level of 0.05 is 2.03, which is larger than the *F* value of $H5_0$ (0.597). The *p*-value is 0.759 and is larger than the significant level (0.05). It suggests that $H5_0$ is accepted and there are no significant interactions between the relevant regression equations and the search engines in terms of R^2.

The results of these hypothesis tests imply that the performances of the regression models are consistent in the two search engines.

4.2.3. Discussion

Surprisingly, the most commonly-used regression model (Equation 7) did not outperform the other regression models in the study. In past studies, Equation (7) was widely used to describe the descending relevance trends of search results returned from search engines. However, in this study, R^2 from the corresponding regression analysis was 0.3723, which placed Equation (7) at the 6th position among the 8 regression models. This interesting finding justifies the importance of this study and suggests

Table 8. The Results for the Follow-up Tukey Test for $H1_0$

(I) Regression Model		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	*2*	-.1624*	.01443	.000	-.2064	-.1184
	3	-.2780*	.01443	.000	-.3220	-.2340
	4	-.3201*	.01443	.000	-.3641	-.2761
	5	.0974*	.01443	.000	.0534	.1414
	6	.0974*	.01443	.000	.0534	.1414
	7	.0974*	.01443	.000	.0534	.1414
	8	.2178*	.01443	.000	.1738	.2618
2	*1*	.1624*	.01443	.000	.1184	.2064
	3	-.1156*	.01443	.000	-.1596	-.0716
	4	-.1577*	.01443	.000	-.2017	-.1137
	5	.2599*	.01443	.000	.2158	.3039
	6	.2599*	.01443	.000	.2158	.3039
	7	.2599*	.01443	.000	.2158	.3039
	8	.3802*	.01443	.000	.3362	.4242
3	*1*	.2780*	.01443	.000	.2340	.3220
	2	.1156*	.01443	.000	.0716	.1596
	4	-.0421	.01443	.072	-.0861	.0019
	5	.3754*	.01443	.000	.3314	.4194
	6	.3754*	.01443	.000	.3314	.4194
	7	.3754*	.01443	.000	.3314	.4194
	8	.4958*	.01443	.000	.4518	.5398
4	*1*	.3201*	.01443	.000	.2761	.3641
	2	.1577*	.01443	.000	.1137	.2017
	3	.0421	.01443	.072	-.0019	.0861
	5	.4175*	.01443	.000	.3735	.4615
	6	.4175*	.01443	.000	.3735	.4615
	7	.4175*	.01443	.000	.3735	.4615
	8	.5379*	.01443	.000	.4939	.5819
5	*1*	-.0974*	.01443	.000	-.1414	-.0534
	2	-.2599*	.01443	.000	-.3039	-.2158
	3	-.3754*	.01443	.000	-.4194	-.3314
	4	-.4175*	.01443	.000	-.4615	-.3735
	6	0.0000	.01443	1.000	-.0440	.0440
	7	0.0000	.01443	1.000	-.0440	.0440
	8	.1204*	.01443	.000	.0764	.1644
6	*1*	-.0974*	.01443	.000	-.1414	-.0534
	2	-.2599*	.01443	.000	-.3039	-.2158
	3	-.3754*	.01443	.000	-.4194	-.3314
	4	-.4175*	.01443	.000	-.4615	-.3735
	5	0.0000	.01443	1.000	-.0440	.0440
	7	0.0000	.01443	1.000	-.0440	.0440
	8	.1204*	.01443	.000	.0764	.1644
7	*1*	-.0974*	.01443	.000	-.1414	-.0534
	2	-.2599*	.01443	.000	-.3039	-.2158
	3	-.3754*	.01443	.000	-.4194	-.3314
	4	-.4175*	.01443	.000	-.4615	-.3735
	5	0.0000	.01443	1.000	-.0440	.0440
	6	0.0000	.01443	1.000	-.0440	.0440
	8	.1204*	.01443	.000	.0764	.1644
8	*1*	-.2178*	.01443	.000	-.2618	-.1738
	2	-.3802*	.01443	.000	-.4242	-.3362
	3	-.4958*	.01443	.000	-.5398	-.4518
	4	-.5379*	.01443	.000	-.5819	-.4939
	5	-.1204*	.01443	.000	-.1644	-.0764
	6	-.1204*	.01443	.000	-.1644	-.0764
	7	-.1204*	.01443	.000	-.1644	-.0764

Table 9. The results for the follow-up Tukey test for $H2_0$

(I) Domain Category		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	*2*	-.0324*	.01020	.009	-.0588	-.0061
	3	.0200	.01020	.205	-.0063	.0464
	4	-.0098	.01020	.771	-.0362	.0165
2	*1*	.0324*	.01020	.009	.0061	.0588
	3	.0524*	.01020	.000	.0261	.0788
	4	.0226	.01020	.121	-.0037	.0490
3	*1*	-.0200	.01020	.205	-.0464	.0063
	2	-.0524*	.01020	.000	-.0788	-.0261
	4	-.0298*	.01020	.019	-.0562	-.0035
4	*1*	.0098	.01020	.771	-.0165	.0362
	2	-.0226	.01020	.121	-.0490	.0037
	3	.0298*	.01020	.019	.0035	.0562

Table 10. The results for $H4_0$ and $H5_0$

Factor	Type III Sum of Squares	df	F	Sig.
Regression model	13.162	7	357.285	0.000
Search engine	0.000	1	0.086	0.077
Interactions	0.022	7	0.597	0.759

that a more appropriate regression model could replace Equation (7) to achieve more effective ranking evaluation.

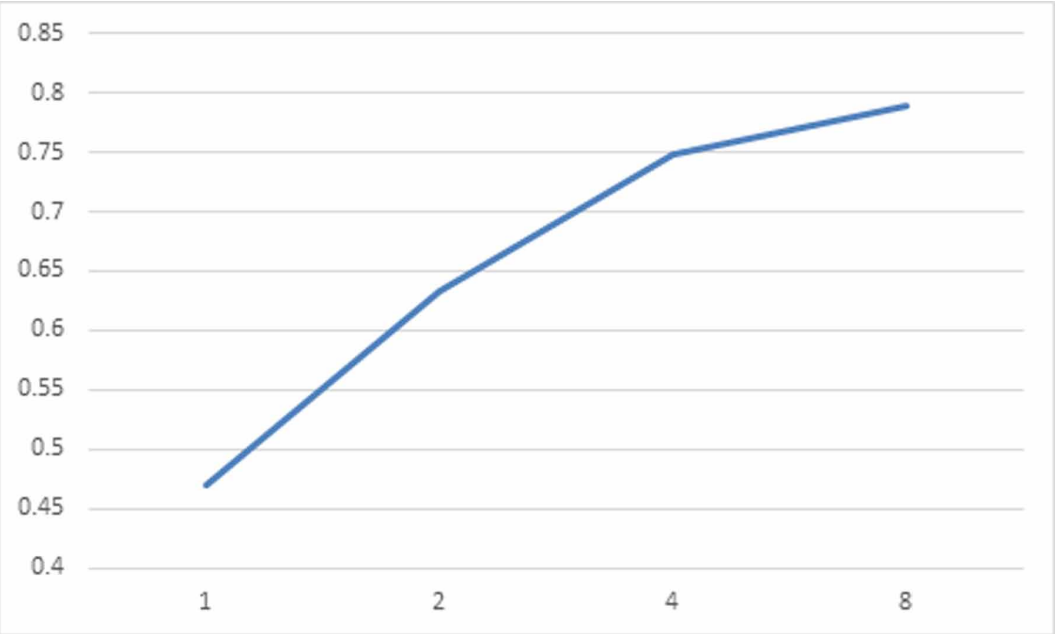
Among the 8 regression models tested, the performance of Model 4 was the best. The corresponding R^2 in the regression analysis was 0.9471 according to the *ANOVA* and follow-up *Tukey* test results. Therefore, Model 4 as the most appropriate regression model is recommended in describing the relevance-descending trend of items returned from search engines.

In the set of the tested regression models, regression models described by Equations (1), (2), (3), and (4) can be classified as one group because the parameters for *b1* in the equations are similar $(\frac{1}{X^1}, \frac{1}{X^{1/2}}, \frac{1}{X^{1/4}}, \frac{1}{X^{1/8}})$. If *t* in Equation (9) below is defined as a variable, the effect of *t* on R^2 of these regression models can be observed in Figure 7:

$$RM(X,t) = b0 + \frac{b1}{X^{1/t}} \tag{9}$$

In Figure 7, the *X*-axis represents variable *t* while the *Y*-axis represents R^2. The relationship between variable *t* in Equation (9) and R^2 are illustrated in Figure 7, and it is quite clear that as the

Figure 7. The Effect of t on R^2 of the Regression Models



value of *t* increases, the corresponding R^2 increases. The range of R^2 is 0.3201. As we know, the higher the R^2, the better the corresponding regression model represents the relevance-decreasing trend of search results from a search engine. It is interesting that when *t* is 3, the increase of R^2 is not noticeable, which is confirmed by Table 10. In Table 10, the difference between Equation (4) and Equation (1) and the difference between Equation (4) and Equation (2) are statistically significant, but the difference between Equation (4) and Equation (1) is not statistically significant.

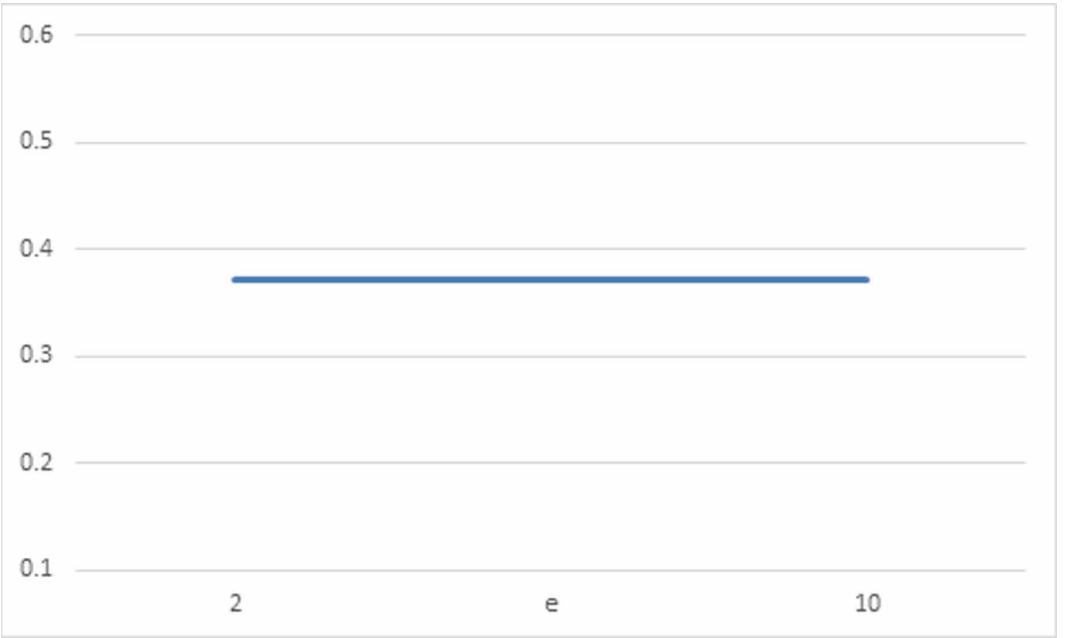
In the set of the tested regression models, regression models represented by Equations (5), (6), and (7) can be classified as one group because the parameters for *b1* in the equations are similar $\left(\frac{1}{Log_2(X)}, \frac{1}{Log_e(X)}, \frac{1}{Log_{10}(X)}\right)$. If *t* is defined as a variable in Equation (10) below, the effect of *t* on R^2 of these regression models can be observed in Figure 8:

$$RM(X,t) = b0 + \frac{b1}{Log_t(X)} \tag{10}$$

In Figure 8, the *X*-axis represents variable *t* while the *Y*-axis represents R^2. The relationship between variable *t* in Equation (10) and R^2 is displayed in Figure 8. Notice that as the value of *t* increases, the corresponding R^2 almost remains the same. The range of R^2 is 0. In other words, the change in variable *t* has little impact on the corresponding R^2.

In this study, 4 different subject categories (*Health, News and Media, Science and Technology,* and *Economy and Business*) were defined. It was important to compare them in terms of the regression model performance. It turned out that *News and Media* (0.5279), *Economy and Business* (0.5053), *Health* (0.4955), and *Science and Technology* (0.4755) ranked first, second, third, and fourth, respectively. Figure 9 shows the results. In Figure 9, the *X*-axis represents 4 different domain

Figure 8. The Effect of t on R^2 of the Regression Models



categories while the *Y*-axis represents R^2. It seems that these regression models worked better in the subject category of *News and Media*.

In this study, two search engines (Google and Bing) were employed. It was important to compare them also in terms of regression model performance. The R^2 score mean of the regression models for Google is 0.5731, and for Bing 0.5727. The mean difference is only 0.004. As depicted in Figure 6, there are no significant differences between the two search engines. It suggests that the performances of the tested regression models are consistent with the search engines employed.

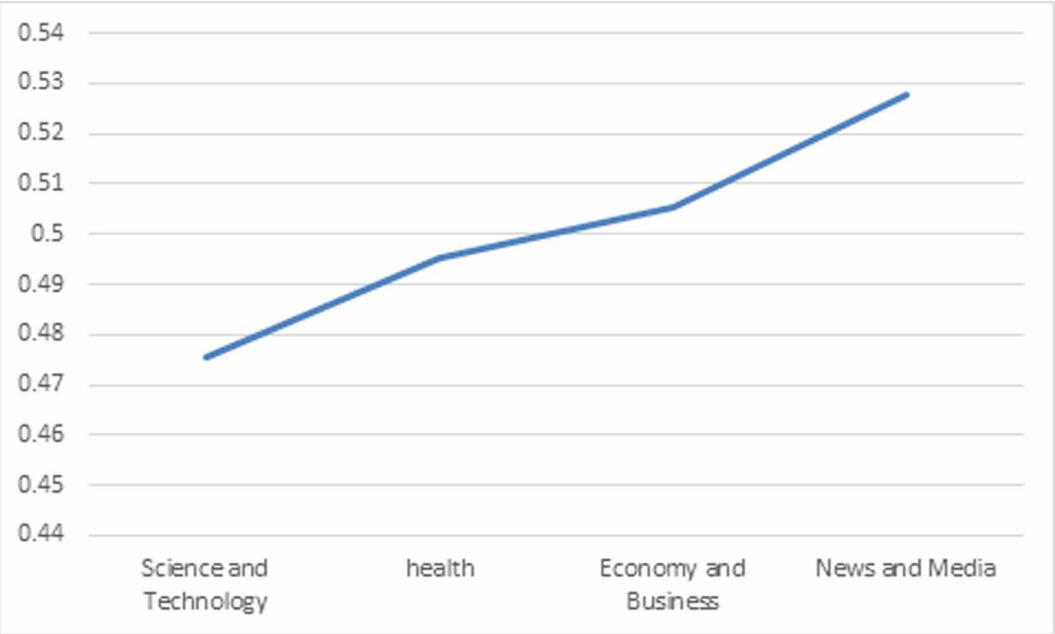
Among the 8 regression models, Model 8 (or Equation 8) had the worst performance. To further confirm this finding, two extra models similar to Regression Model 8 were added, and corresponding tests were conducted. The following are the added equations (Equation 11 and Equation 12):

$$RM9(X) = b0 + \frac{b1}{4^X} \tag{11}$$

$$RM10(X) = b0 + \frac{b1}{8^X} \tag{12}$$

If *t* is defined as a variable in the following equation (Equation 13), the effect of *t (t=2, 4,* and *8)* on R^2 of these three derived regression models can be observed in Figure 10. In Figure 10, *X*-axis represents variable *t* while the *Y*-axis represents R^2. It is quite clear that as *t* increases, the corresponding R^2 decreases dramatically.

Figure 9. The Effect of t on R^2 of the Subject categories



$$RM(X,t) = b0 + \frac{b1}{t^X} \tag{13}$$

5. CONCLUSION

As Internet technology continues to change and hopefully improve lives and societies worldwide, effective global information management becomes increasingly critical, and effective Internet information retrieval systems become more and more significant in providing Internet users worldwide with accurate and complete information.

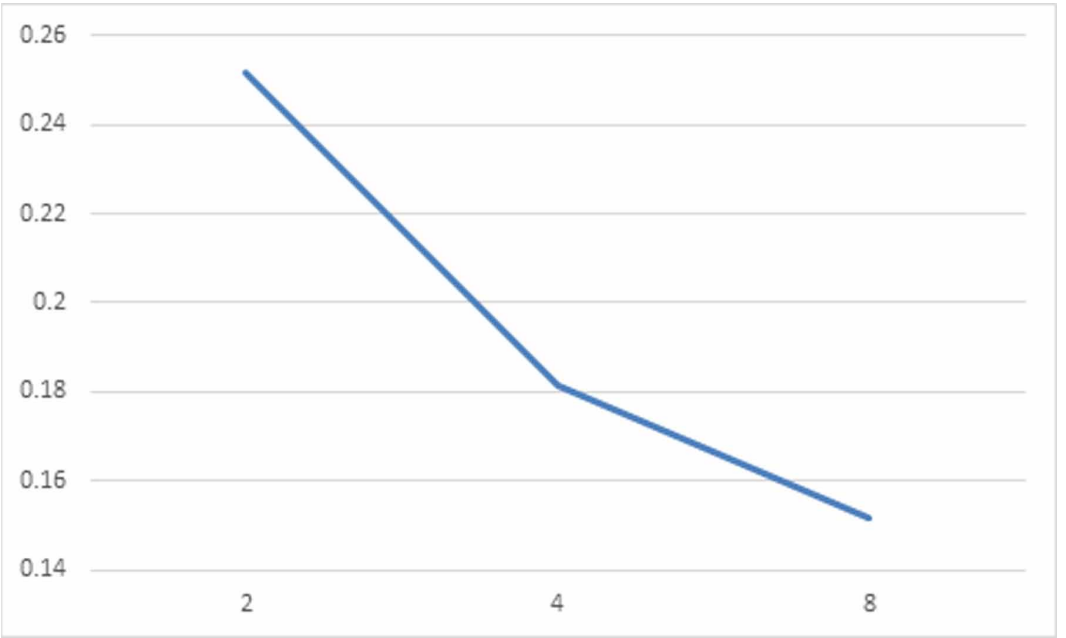
Both average Internet users and business intelligence systems seem to accept the search results presented by the search engines as a list of decreasing relevance, while in reality it might not be so. Users tend to browse only the first 20-30 items risking missing important information because they might not have been properly positioned on the result list by the search engine's ranking algorithms. Therefore, accurate relevance ranking of search result items as returned by search engines is extremely important.

This study aimed to discover the effective measurement of search results from search engines and find the best regression model in describing the downward relevance trends of the returned result lists.

To help achieve the purpose, 5 hypotheses were proposed to explore relationships among search engines, regression models, subject categories, and their interactions. Two search engines, Google and Bing, were employed in the study. Six search tasks from each of the four subject categories (*Health, News and Media, Science and Technology,* and *Economy and Business*) were designed and submitted to each search engine. The returned data sets were randomly assigned to 32 evaluators for independent relevance evaluations. Relevance scores were plugged into 8 regression models. Consequently, 384 R^2 from the regression models were collected.

Based on the R^2 collected, two two-factor *ANOVA* tests were conducted to test the proposed null hypotheses. Hypotheses $H1_0$ and $H2_0$ were rejected and Hypotheses $H3_0$, $H4_0$, and $H5_0$ were accepted. Significant differences were discovered among regression models and among subject categories, but

Figure 10. The Effect of t on R^2 of the Regression Models



no significant interactions were found between them in terms of R^2. No significant differences were found between the two search engines, and no significant interactions existed between regression models and search engines in terms of R^2.

Equation (4) as shown below was identified to be the best fit regression model among the 8 regression models proposed across all 4 subject categories and for both search engines.

$$RM4(X) = b0 + \frac{b1}{X^{1/8}} \tag{14}$$

The findings of this study have both theoretical and practical implications. People generally assume that relevance of items on a results list from a search engine follows a downward trend; however, there was no appropriate model to describe the trend. The result of this study presents a best model to describe the trend which would make relevance evaluation and ranking of a retrieval results list more sound and plausible. This model can also be applied to OPAC and other information retrieval systems.

Several limitations of this study were recognized. First, the number of returned items from each search task was limited to 50. Although previous studies showed that most people would only read the first page of the returned results, a longer result list would be better for illustrating the relevance-decreasing pattern. Secondly, although this study showed that Google and Bing had similar relevance-descending patterns in terms of R^2 from regression analyses, it would have been better to include more search engines in this study. Finally, if more regression models had been proposed and tested, the findings of this study would be even more convincing.

Despite these limitations, the findings of this study can be applied to relevance evaluation of search results from other information retrieval systems such as OPAC, can help make search engine evaluations more accurate and sound, and can provide global information management personnel with valuable insights.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under grant no. 71420107026.

REFERENCES

Al-Maskari, A., Sanderson, M., & Clough, P. (2007). The Relationship between IR Effectiveness Measures and User Satisfaction. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 773-774). ACM. doi:10.1145/1277741.1277902

Bar-Ilan, J. (1999). Search Engine Results over Time: A Case Study on Search Engine Stability. *Cybermetrics (Madrid)*, *2*(3).

Bar-Ilan, J. (2004). Search Engine Ability to Cope with the Changing Web. In M. Levene & A. Poulovassilis (Eds.), *Web Dynamics* (pp. 195–215). Springer Berlin Heidelberg. doi:10.1007/978-3-662-10874-1_9

Bar-Ilan, J., Keenoy, K., Yaari, E., & Levene, M. (2007). User Rankings of Search Engine Results. *Journal of the American Society for Information Science and Technology*, *58*(9), 1254–1266. doi:10.1002/asi.20608

Barysevich, A. (2017). 2017's Four Most Important Ranking Factors, According to SEO Industry Studies. Retrieved 2017-06-01 from https://www.searchenginejournal.com

Bilal, D. (2002). Children's Use of the Yahooligans! Web Search Engine. III. Cognitive and Physical Behaviors on Fully Self-Generated Search Tasks. *Journal of the American Society for Information Science and Technology*, *53*(13), 1170–1183. doi:10.1002/asi.10145

Bing. (n.d.). Retrieved on 2016-01-15 from http://www.bing.com

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Brin, S., & Page, L. (2012). Reprint of: The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, *56*(18), 3825–3833. doi:10.1016/j.comnet.2012.10.007

Caprio, D. D., Santos-Arteaga, F. J., & Tavana, M. (2015). Technology Development through Knowledge Assimilation and Innovation: A European Perspective. *Journal of Global Information Management*, *23*(2), 48–93. doi:10.4018/JGIM.2015040103

Carterette, B., & Jones, R. (2008). Evaluating Search Engines by Modeling the Relationship between Relevance and Clicks. *In Proceedings of Twenty-First Annual Conference on Neural Information Processing Systems* (pp. 217-224).

Chatterjee, S., Kar, A. K., & Gupta, M. P. (2017). Critical Success Factors to Establish 5G Network in Smart Cities: Inputs for Security and Privacy. *Journal of Global Information Management*, *25*(2), 15–37. doi:10.4018/JGIM.2017040102

Chignell, M. H., Gwizdka, J., & Bodner, R. C. (1999). Discriminating Meta-Search: A Framework for Evaluation. *Information Processing & Management*, *35*(3), 337–362. doi:10.1016/S0306-4573(98)00065-X

Chu, H., & Rosenthal, M. (1996). Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology. In *Proceedings of the Annual Meeting-American Society for Information Science* (Vol. 33, pp. 127-135).

comScore. (2015). comScore Releases November 2015 U.S. Search Engine Rankings. Retrieved 2015-12-15 from http://www.comscore.com/Insights/Market-Rankings/comScore-Releases-November-2015-US-Desktop-Search-Engine-Rankings

da Costa Pereira, C., Dragoni, M., & Pasi, G. (2012). Multidimensional Relevance: Prioritized Aggregation in a Personalized Information Retrieval Setting. *Information Processing & Management*, *48*(2), 340–357. doi:10.1016/j.ipm.2011.07.001

Davis, M. W. (1996). New Experiments In Cross-Language Text Retrieval At NMSU's Computing Research Lab. In *Proceedings of TREC*-5, Gaithersburg, MD.

Dean, B. (2016). Google's 200 Ranking Factors: The Complete List. Retrieved 2017-06-01 from http://backlinko.com

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Efron, M., & Golovchinsky, G. (2011). Estimation Methods for Ranking Recent Information. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 495-504). ACM.

Gey, F. C., Kando, N., & Peters, C. (2005). Cross-Language Information Retrieval: The Way Ahead. *Information Processing & Management*, *41*(3), 415–431. doi:10.1016/j.ipm.2004.06.006

Google. (2016). Retrieved 2016-01-15 from http://www.google.com

Hassan, I., & Zhang, J. (2001). Image Search Engine Feature Analysis. *Online Information Review*, *25*(2), 103–114. doi:10.1108/14684520110390042

Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 50-57). ACM.

Hung, S. Y., Huang, W. M., Yen, D. C., Chang, S. I., & Lu, C. C. (2016). Effect of Information Service Competence and Contextual Factors on the Effectiveness of Strategic Information Systems Planning in Hospitals. *Journal of Global Information Management*, *24*(1), 14–36. doi:10.4018/JGIM.2016010102

Internet live stats. (2015). The Summary of Internet Users in 2014. Retrieved 2016-1-12 from http://www.internetlivestats.com/internet-users/

Jacsó, P. (2008). Savvy Searching How Many Web-wide Search Engines Do We Need? *Online Information Review*, *32*(6), 860–865. doi:10.1108/14684520810923971

Jansen, B. J., & Pooch, U. (2001). A Review of Web Searching Studies and a Framework for Future Research. *Journal of the American Society for Information Science and Technology*, *52*(3), 235–246. doi:10.1002/1097-4571(2000)9999:9999<::AID-ASI1607>3.0.CO;2-F

Jansen, B. J., & Spink, A. (2006). How Are We Searching the World Wide Web? A Comparison of Nine Search Engine Transaction Logs. *Information Processing & Management*, *42*(1), 248–263. doi:10.1016/j.ipm.2004.10.007

Järvelin, K., & Kekäläinen, J. (2002). Cumulated Gain-Based Evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446. doi:10.1145/582415.582418

Kaufmann, E., & Bernstein, A. (2010). Evaluating the Usability of Natural Language Query Languages and Interfaces to Semantic Web Knowledge Bases. *Journal of Web Semantics*, *8*(4), 377–393. doi:10.1016/j.websem.2010.06.001

Khatwani, G., & Srivastava, P. R. (2017). An Optimization Model for Mapping Organization and Consumer Preferences for Internet Information Channels. *Journal of Global Information Management*, *25*(2), 88–115. doi:10.4018/JGIM.2017040106

Lane, V. R., Khuntia, J., Parthasarathy, M., & Hazarika, B. B. (2017, July). The Impact of the Internet on Values in India: Shifts in Self-Enhancement and Self-Transcendence Amongst Indian Youth. *Journal of Global Information Management*, *25*(3), 98–120. doi:10.4018/JGIM.2017070106

Lee, H., Choi, J., Kim, K. K., & Lee, A. R. (2014). Impact of Anonymity on Information Sharing through Internal Psychological Processes: A Case of South Korean Online Communities. *Journal of Global Information Management*, *22*(3), 57–77. doi:10.4018/jgim.2014070103

Lin, L., Xu, Z., Ding, Y., & Liu, X. (2013). Finding Topic-Level Experts in Scholarly Networks. *Scientometrics*, *97*(3), 797–819. doi:10.1007/s11192-013-0988-6

Purcell, K., Brenner, J., & Rainie, L. (2012) Search Engine Use 2012. Retrieved 2016-1-12 from http://www.pewinternet.org/2012/03/09/search-engine-use-2012/

Roztocki, N., & Weistroffer, H. R. (2011). Information Technology Success Factors and Models in Developing and Emerging Economies. *Information Technology for Development*, *17*(3), 163–167. doi:10.1080/02681102.2011.568220

Silic, M., & Back, A. (2016). What Are the Keys to a Successful Mobile Payment System? Case of Cytizi: Mobile Payment System. *Journal of Global Information Management*, *24*(3), 1–20. doi:10.4018/JGIM.2016070101

Soja, P. (2016). Reexamining Critical Success Factors for Enterprise System Adoption in Transition Economies: Learning from Polish Adopters. *Information Technology for Development*, *22*(2), 279–305. doi:10.1080/02681102.2015.1075189

Spink, A. & Jansen, B. J. (2004). A Study of Web Search Trends. *Webology*, *1*(2).

Teo, T. S. H. (2007). Organizational Characteristics, Modes of Internet Adoption and Their Impact: A Singapore Perspective. *Journal of Global Information Management*, *15*(2), 91–117. doi:10.4018/jgim.2007040104

Thelwall, M. (2008). Quantitative Comparisons of Search Engine Results. *Journal of the American Society for Information Science and Technology*, *59*(11), 1702–1710. doi:10.1002/asi.20834

Tominski, C., Abello, J., & Schumann, H. (2009). CGV-- An Interactive Graph Visualization System. *Computers & Graphics*, *33*(6), 660–678. doi:10.1016/j.cag.2009.06.002

Vaughan, L. (2004). New Measurements for Search Engine Evaluation Proposed and Tested. *Information Processing & Management*, *40*(4), 677–691. doi:10.1016/S0306-4573(03)00043-8

Vaughan, L., & Thelwall, M. (2004). Search Engine Coverage Bias: Evidence and Possible Causes. *Information Processing & Management*, *40*(4), 693–707. doi:10.1016/S0306-4573(03)00063-3

Wilbur, W. J., & Sirotkin, K. (1992). The Automatic Identification of Stop Words. *Journal of Information Science*, *18*(1), 45–55. doi:10.1177/016555159201800106

Wu, I. C. (2011). Toward Supporting Information-Seeking and Retrieval Activities Based on Evolving Topic-Needs. *The Journal of Documentation*, *67*(3), 525–561. doi:10.1108/00220411111124578

Xie, H. I. (2000). Shifts of Interactive Intentions and Information-Seeking Strategies in Interactive Information Retrieval. *Journal of the American Society for Information Science*, *51*(9), 841–857. doi:10.1002/(SICI)1097-4571(2000)51:9<841::AID-ASI70>3.0.CO;2-0

Yahoo. (2016). Retrieved 2016-1-15 from http://www.Yahoo.com

Zhang, J., & Dimitroff, A. (2005). The Impact of Webpage Content Characteristics on Webpage Visibility in Search Engine Results (Part I). *Information Processing & Management*, *41*(3), 665–690. doi:10.1016/j.ipm.2003.12.001

Zhang, J., Fei, W., & Le, T. (2010). Evaluation of Six Google Search Features. *Online (Bergheim)*, *34*(6), 24–28.

Zhang, J., Fei, W., & Le, T. (2013). A Comparative Analysis of the Search Feature Effectiveness of the Major English and Chinese Search Engines. *Online Information Review*, *37*(2), 217–230. doi:10.1108/OIR-07-2011-0099

Zhang, J., & Korfhage, R. R. (1999). A Distance and Angle Similarity Measure Method. *Journal of the Association for Information Science and Technology*, *50*(9), 772.

Zhang, J., & Lin, S. (2007). Multiple Language Supports in Search Engines. *Online Information Review*, *31*(4), 516–532. doi:10.1108/14684520710780458

Zheng, Z., Chen, K., Sun, G., & Zha, H. (2007). A Regression Framework for Learning Ranking Functions Using Relative Relevance Judgments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 287-294). ACM. doi:10.1145/1277741.1277792

Zhou, B., & Yao, Y. (2010). Evaluating Information Retrieval System Performance Based on User Preference. *Journal of Intelligent Information Systems*, *34*(3), 227–248. doi:10.1007/s10844-009-0096-5

Jin Zhang is a Tenured full professor in the School of Information Studies at the University of Wisconsin Milwaukee

Xin Cai is a Ph.D. candidate in the School of Information Science, University of Wisconsin-Milwaukee, USA. He holds a Master's degree in information science from Central China Normal University, China, and a Bachelor's degree in computer science from Shenyang Normal University, China. His research interests include information retrieval and system, data mining, and domain analysis.

Taowen Le is a Tenured full professor and former department chair of information systems & technologies at Weber State University in Utah, U.S.A. Currently the Business Division Chair of Utah Academy of Sciences, Arts, and Letters.

Wei Fei is a Doctor of Library Science and is currently the Assistant curator of Suzhou Library.

Feicheng Ma is a Full professor and former dean of the School of Information Management at Wuhan University, China.