

BIROn - Birkbeck Institutional Research Online

Zhitomirsky-Geffet, M. and Bari-Ilan, J. and Levene, Mark (2016) Testing the stability of “wisdom of crowds” judgments of search results over time and their similarity with the search engine rankings. *ASLIB Journal of Information Management* 68 (4), pp. 407-427. ISSN 2050-3806.

Downloaded from: <https://eprints.bbk.ac.uk/id/eprint/14848/>

Usage Guidelines:

Please refer to usage guidelines at <https://eprints.bbk.ac.uk/policies.html>
contact lib-eprints@bbk.ac.uk.

or alternatively



Testing the stability of “wisdom of crowds” judgments of search results over time and their similarity with the search engine rankings

Journal:	<i>Aslib Journal of Information Management</i>
Manuscript ID	AJIM-10-2015-0165.R3
Manuscript Type:	Research Paper
Keywords:	ranking, relevance judgment, wisdom of crowds, change in time, user evaluation of search results, change coefficient

SCHOLARONE™
Manuscripts

Review

Testing the stability of “wisdom of crowds” judgments of search results over time and their similarity with search engine rankings

Maayan Zhitomirsky-Geffet
Bar-Ilan University
Ramat-Gan, Israel
Maayan.Zhitomirsky-Geffet@biu.ac.il

Judit Bar-Ilan
Bar-Ilan University
Ramat-Gan, Israel
Judit.Bar-Ilan@biu.ac.il

Mark Levene
Birkbeck University
London, UK
Mark@dcs.bbk.ac.uk

Corresponding author: Maayan Zhitomirsky-Geffet, e-mail: maayan.zhitomirsky-geffet@biu.ac.il.

Abstract

Purpose: One of the under-explored aspects in the process of user information seeking behaviour is influence of time on relevance evaluation. It has been shown in previous studies that individual users might change their assessment of search results over time. It is also known that aggregated judgments of multiple individual users can lead to correct and reliable decisions; this phenomenon is known as the “wisdom of crowds”. The aim of this study is to examine whether aggregated judgments will be more stable and thus more reliable over time than individual user judgments.

Design/Methods: In this study two simple measures are proposed to calculate the aggregated judgments of search results and compare their reliability and stability to individual user judgments. In addition, the aggregated “wisdom of crowds” judgments were used as a means to compare the differences between human assessments of search results and search engine’s rankings. A large-scale user study was conducted with 87 participants who evaluated two different queries and four diverse result sets twice, with an interval of two months. Two types of judgments were considered in this study: 1) relevance on a 4-point scale, and 2) ranking on a 10-point scale without ties.

Findings: It was found that aggregated judgments are much more stable than individual user judgments, yet they are quite different from search engine rankings.

Practical implications: The proposed “wisdom of crowds” based approach provides a reliable reference point for the evaluation of search engines. This is also important for exploring the need of personalization and adapting search engine’s ranking over time to changes in users preferences.

Originality/Value: This is a first study that applies the notion of “wisdom of crowds” to examine the under-explored phenomenon in the literature of “change in time” in user evaluation of relevance.

Keywords: ranking, relevance judgment, change in time, wisdom of crowds

Research paper

1. Introduction

Numerous general models of information seeking and web searching behaviour have been proposed in the past such as (Ellis, 1989; Bates, 1989; Kuhlthau, 1991; Dervin, 1992; Johnson and Meishke, 1993; Marchionini, 1995; Spink, 1997; Wilson, 1999; Fisher et al., 2005; Knight and Spink, 2008; Du and Spink, 2010; Case, 2012). Relevance is a central notion in information science and is an important part of user information seeking models (Saracevic, 2007). This study investigates an under-explored topic in the literature (Saracevic, 2007): stability and change of user assessment of search results over time. Human evaluation of documents relevance is a complex process that requires coordination of multiple cognitive tasks (Du and Spink, 2011). User result evaluation is needed in many fields and has many purposes, hence it is important to understand the factors and phenomena behind it. This study aims to extend the understanding of the result evaluation component of the proposed web search behavior models, with respect to the temporal change factor. In this broad context, this research contributes to modelling the change in user relevance evaluation behaviour over time.

As stated by (Saracevic, 2007, p. 2139): “The role of research is to make relevance complexity more comprehensible formally and possibly even more predictable”. Accurate ranking of search results according to the users’ preferences is one of the most important challenges of the modern search systems. However, previous research found a low correlation between users’ and search engines’ rankings of search results (Vaughan, 2004; Veronis, 2006; Bar-Ilan, Keenoy, Yaari and Levene, 2007; Lewandowski, 2008; Bar-Ilan and Levene, 2011), thus, leading to a conclusion that more work is required to improve the systems’ ability to assess documents’ relevance.

Previous work above and those reviewed in (Saracevic, 2007) concentrated on the successive or evolving search processes, where further iterations are used to refine and improve the search. It is known that users’ information needs, evaluation criteria and preference of results, as well as query formulation and retrieved result sets tend to change during the search process, since users better understand their needs at the end of the process rather than at the beginning, and try to refine their search to get the optimal results. As opposed to the above works investigating an evolution of search and result evaluation process, this study explores a different dimension of change in user evaluation of relevance: the “change in time”. This change, if it exists, reflects the essential subjectivity and instability of user perception and evaluation of

relevance, and thus might reveal the inherent complexity, subjectivity and vagueness/fuzziness in users' perception of relevance. This type of change might be discovered when other factors of influence are neutralised (i.e. in independent evaluation sessions with identical tasks, environments, goals and data but at two different points in time). In other words, if users were asked to choose a relevance grade or a rank for each result, given the same query and result set, would these assessments remain stable over time? Would users provide similar relevance judgments and ranks to the same results and queries in a few weeks or months?

It was shown in previous work (Scholer et al., 2011) that individual users might change their assessments of search results over time due to subjectivity in human relevance perception or even human error. Inter-user agreement on ranking of search results has also been shown to be quite low due to subjectivity in human judgments (Bar-Ilan et al., 2007; Bar-Ilan and Levene, 2011). On the other hand, it is also known that in many fields of knowledge aggregated judgments of multiple individual users lead to more correct and reliable decisions; this phenomenon is called "wisdom of crowds" (Cooper et al., 2010; Giles, 2005; Surowiecki, 2005; Preis et al., 2013; Harshavardhan et al., 2012; Bollen and Mao, 2011; Mortensen et al., 2014; Zhitomirsky-Geffet and Erez, 2014; Cen et al., 2009; Bao et al., 2007). Therefore, this study's research goal is to examine the level of change and stability of aggregated judgments compared to individual user judgments. Accordingly, the two main research questions tested in this study are:

- 1) Whether and how aggregated judgments will be more stable and thus more reliable over time than individual user judgments?
- 2) Whether and to what extent are the search engines' ranking similar/different from the "wisdom of crowds" ranking?

As noted above, previous research reveals quite a high level of disagreement between the ranking of search engines and rankings produced by individual. In this context an additional goal of this study is to examine to what extent the aggregated "wisdom of crowds" judgments differ from the ranking of search engines.

The rest of this paper is organized as follows. In the next section the related work is reviewed. Then our study setup is described, and following that the results are presented and discussed. Finally, some conclusions and future research directions are provided.

2. Related work

First, a review of some previous studies is presented, which apply "wisdom of crowds"-based techniques to improve and learn search result relevance and ranking. Then, the most relevant user studies are

reviewed, which are related to agreement on ranking and relevance judgments and comparison between user and search engine ranking.

2.1 “Wisdom of crowds” techniques and information retrieval

In recent years, a number of articles have suggested using social tags as a source of “wisdom of crowds” for improving ranking of search results (e.g., Yanbe *et al.*, 2007; Bao *et al.*, 2007; Zhang *et al.*, 2009; Choochaiwattana and Spring, 2009; Zhitomirsky-Geffet and Daya, 2015). Yanbe *et al.* (2007) suggested enhancing result ranking by integrating the PageRank algorithm with the tag information on social bookmarking sites. Bao *et al.* (2007) devised two algorithms for ranking according to social bookmarking: 1) the SocialSimRank algorithm which assesses the resemblance between the query and the tags; and 2) the SocialPageRank algorithm which measures the quality of a page according to its popularity. Their study indicates that these two algorithms significantly improved the quality of result ranking. A similar method was presented in an additional study (Zhang *et al.*, 2009). This method ranks search results according to a query's resemblance to the tags, with the rank weight determined by the popularity of the tags. Another study (Choochaiwattana and Spring, 2009) considered the number of social tags that matched the query terms. The authors reported that the ranking method that yielded the best results, ranked the document according to the number of users who tagged it on Delicious with tags that matched the terms of the search query. Kawase *et al.* (2014) employed Wikipedia categories constructed by wisdom of crowds as a basis for fingerprints creation for different web services (e.g. Twitter, Flickr, Delicious). The topic coverage of these services' represented by their fingerprints was comparatively analysed. These fingerprints were also shown to be effectively used for a movie recommendation task in the crowdsourcing experiment. Singh *et al.* (2013) developed an eBook recommender system based on content analysis and various social web eResources, e.g. YouTube, Slidershare, Twitter and LinkedIn. Zhitomirsky-Geffet and Daya (2015) presented a technique for using social tags to extract diverse subtopics for a query, and reduction and re-ranking of search results, according to the most prominent and discriminative subtopics.

Another group of investigations used user click-through data as a source of “wisdom of crowds” to infer user relevance preferences of search results (Cen *et al.*, 2009; Agichtein *et al.*, 2006; Dou *et al.*, 2008). Cen *et al.* (2009) showed that it is possible to accurately evaluate relevance of search results based on aggregated click-through information from query logs. The underlying assumption was that a result with a larger amount of clicks is more relevant to the query than a result with fewer clicks. Agichtein *et al.* (2006) proposed an idea of aggregating information from many unreliable user search sessions, instead of treating each user as a reliable expert to predict user relevance assessment of search results. Dou *et al.* (2008) used aggregate click-through logs to learn the ranking of search results, and found that the aggregation of a large number of user clicks is indicative of relevance preferences. Harris (2014) found

that crowds are able to predict the consensus ranking of search results with significantly higher recall when asked to judge document relevance based on their estimate of the consensus decision than when the judgment is based on their personal viewpoint. Zhitomirsky-Geffet et al. (2016) applied a similar methodology for classification of diet ontology’s statements by crowdsourcing. They found that crowds are able to correctly distinguish between consensual and controversial statements when asked to predict the experts’ opinion.

In summary, the above studies demonstrated that “wisdom of crowds”-based techniques applied to various types of user data can increase the reliability of this data for learning relevance preferences and ranking of search results. The goal of the current study is to test whether such techniques of aggregation of user-produced data might increase the stability of user evaluation of search results over time.

2.2 Relevance evaluation and ranking of search results by users and search engines

Lewandowski (2008) conducted a user study with 40 subjects who judged relevance (on a binary scale) of top-20 results of five search engines. He reported quite low precision at 20 results, ranging from 0.37 to 0.52, while Yahoo! and Google outperformed the other search engines and yielded quite similar results. Vaughan (2004) compared 24 subjects’ ranking of four queries’ results with those of Google, AltaVista and Teoma. In his study, Google outperformed the other search engines with 0.72 average correlation between Google’s and subjects’ rankings. Veronis (2006) conducted a user study with 14 students as subjects who judged the relevance of top-10 results of six search engines on 14 topics and 5 queries per topic. He found that Google and Yahoo! significantly outperformed the other search engines but still reached only an average score of 2.3 on a 0-5 relevance scale. A later study examined differences in relevance judgments of results retrieved by Google, Yahoo!, Bing, Yahoo! Kids, and ask Kids search engines for 30 queries formulated by children (Bilal, 2012). Yahoo! and Bing produced a similar percentage in hit overlap with Google (nearly 30%), while Google performed best on natural language queries, and Bing showed a similar precision score ($P=0.69$) on two-word queries. In a recent large-scale study (Lewandowski, 2015) a sample of 1,000 informational and 1,000 navigational queries from a major German search engine was used to compare Google's and Bing's search results. It was found that Google slightly outperformed Bing for informational queries, however, there was a substantial difference between Google and Bing for navigational queries. Google found the correct answer in 95.3% of cases, whereas Bing only found the correct answer 76.6% of the time. These studies did not consider ranking of the results but only compared their relevance grades.

A few studies compared user ranking of search results to popular search engines’ ranking. In a study (Bar-Ilan et al., 2007) users were presented with randomly ordered result sets retrieved from Google, Yahoo!

and MSN (now Bing) and were asked to choose and rank the top-10 results. The findings, generally, showed low similarity between the users and the search engines rankings. In a follow-up study (Bar-Ilan and Levene, 2011), country-specific search results were tested in a similar way. In this case it was shown that at least for Google, the users preferred the results and the rankings of the local Google version over other versions. In (Hariri, 2011) the authors also studied Google rankings and asked whether top results are considered more relevant by the users. In this study the fifth ranked result was judged to be of highest relevance, slightly more than the top ranked result. These studies only asked the users to rank the results, without asking for their relevance judgments.

2.3 Change in time in relevance and ranking evaluation

Saracevic (2007) in his extensive review on relevance discusses the dynamics of relevance evaluation over time, when the information need changes due to information gained during the information search. One of the first studies of dynamic changes was carried out by Rees and Schultz (1967). According to the information retrieval model of Bates (1989) during the iterative process of search the user relevance judgments of the results are influenced by the results of previous search. Later, Spink and Dee (2007) defined a web search model as comprising multiple tasks and cognitive shifts between tasks (e.g. shifts between topic, result evaluation, document, information problem, search strategy). Cognitive shift was defined as a human ability to handle the demands of complex and often multiple tasks resulting from changes due to external forces. Du and Spink (2011) found that evaluation is one of the three most experienced states during multi-tasking search process. Also shifts from one evaluation to another were quite frequent among other shift types. Saracevic (2007) mentions additional studies where the relevance assessments at different points in the information seeking task of more than two participants were investigated (Smithson, 1994; Bruce, 1994; Wang and White, 1995; Bateman, 1998; Vakkari and Hakala, 2000; Vakkari, 2001; Tang and Solomon, 2001). However, the setting of the above mentioned studies is different from the current setting. In the previous studies the users' information need changed as the task evolved. In the current study the participants were explicitly instructed to use the same criteria and goals, the same query and result sets in both rounds of the experiment. The question is what happens in two separate standalone search sessions when the task is identical, and assessed at two different points in time? The only difference between the sessions is that the users saw the given set of documents (or their snippets) once before.

Self-agreement and change in user evaluation of the same search results for the same query is an under-explored area. Scholer et al. (2011) studied repeated relevance judgments of TREC evaluators. They found that quite often (for 15-24% of the documents) the evaluators were not consistent in their decisions, and considered these inconsistencies to be errors made by the assessors. As opposed to their study, here

changes in (ordinary) users' rather than domain experts' judgments are measured, for relevance on a four point-scale as opposed to the binary scale used by them, and also for ranking of the top-ten results.

Scholer et al. (2013) studied the influence of exposure to more or less relevant documents on relevance assessment of documents shown later. They asked their users to evaluate the relevance (on a 4-point scale) 28 documents, where the first three and the last three were identical, thus they saw the same documents for the second time after viewing and judging 25 other documents. In their study the users viewed documents for the second time within the same sessions, while in our study there is a significant gap in time between the two assessments. The reported self-agreement on these three documents was only about 50%. To the best of our knowledge changes in users' rankings over time have not been examined in any previous research.

In summary, it has been shown in the literature that there is a substantial difference between users' and search engines' relevance evaluation of search results. This means that in order to reduce this gap more research is needed into this field. The main differences between the current research and the reviewed literature are as follows. Studies that explored the change in users' search behaviour over time, mostly addressed successive search behaviour and used only one type of evaluation (either ranking or relevance judgment). Neither of them investigated the "wisdom of crowds" evaluation change in time. Conversely, past works that applied "wisdom of crowds" techniques did not explore the temporal factor of information retrieval and evaluation. The most related study by Scholer et al. (2011) used only binary evaluation of documents' relevance and did not check the change in "wisdom of crowds" result evaluation.

3. Method

3.1 Study Setting

3.1.1 Queries

Ideally users choose topics of their interest to search for and make assessments, however when the queries differ between users their judgments cannot be aggregated. Therefore, to test the above research questions two popular scientific topics were selected as queries by the authors, Big Data (in English) and Alzheimer (in Hebrew). In addition to the queries a search scenario was provided as "for the aim of preparing a summary of the topic, based solely on the results in this set" (they did not actually have to submit the summary).

The query topics were not part of the curriculum, and were not studied either in this course or in any other courses the participants took. For each query two separate sets of 20 search results were created. The search results of the first set were collected from Google and Bing, and included top-10 Google and top-

10 Bing, supplemented, because of the partial overlap between the top results of the two search engines. The second set comprised the Google results displayed on the first and the tenth result pages (i.e. results 1 to 10 and 101 to 110). Thus, four different tasks were defined each with a different query and a different result set for each query: AlzheimerGoogle10&100, BigDataGoogle10&100, AlzheimerGoogle&Bing and BigDataGoogle&Bing. Google and Bing are the two leading search engines according to comScore (2015), and this is why we chose to present results from these search engines.

3.1.2 Participants

Two randomly created groups of 42 and 45 Information Science students, who participated in the “Introduction to Information Science” course, were asked to judge the results. No specific demographic data was collected for the purpose of this study. Each group was presented with two out of the above four tasks, one for each query. Every result set was judged by only one of the groups. The order of presentation of the results to the students was random, to avoid prior bias in their judgments. The students were instructed to judge the results in the set with respect to the query with the aim of preparing a summary of the topic, based solely on the results in this set (they did not actually have to submit the summary). Two types of judgments were featured: relevance judgments on a scale of four: not relevant (1), slightly relevant (2), somewhat relevant (3) and relevant (4), where ties were allowed; ranks for the top-10 out of 20 results with no ties allowed. The tasks with queries and results for judgment were presented to the participants in Google forms and included title, URL and snippet as displayed by the search engine for each result along with two types of judgment scales.

Two months later the same participants were asked to judge the same result sets for the same queries with the same evaluation criteria and instructions but presented in a different random order. This time the same set of results was presented in a different random order to prevent the students from copying or fully recalling their first judgments. We note that the first round of evaluation took place about six weeks after the fall semester started, and the second time occurred at the end of the fall semester.

The participants filled-in the Google forms using their personal computers (laptops or desktops) for both rounds. Most of them performed judged the relevance and ranked both queries on the same day in the round, although this was not a requirement.

3.1.3 Relevance judgment and engines’ rankings

The two types of judgments were employed, since they test different cognitive processes executed by the users and we wanted to understand the differences between them. When assessing relevance of a document to a query, this can be done independently from the other retrieved documents and thus requires

a smaller amount of cognitive shifts and their coordination, while for ranking the whole set of retrieved documents must be taken into account engaging a higher level of multi-tasking and coordinated cognitive shifts. The choice of the relevance scale was based on our preliminary experiment, where we asked 27 users to decide on the number of relevance categories and then assign each search result to a category. The average number of categories was 4.1, which led us to the decision in these experiments to use a 4-point scale for relevance. It seems more reasonable to ask the participants rank only top-10 results than all 20, as it would require too much cognitive effort, coordination of multiple cognitive shifts (Du and Spink, 2011), and time to compare and uniquely rank 20 results of two different queries. Also search engines normally present only the top-10 results on the first page, which are considered to be the most important for users (Jansen & Spink, 2006; Chitika, 2013, p.5).

3.2 Presentation bias

In each round all the participants in a given group saw the results in the same order. While, in general, the order of results might have influenced the judgments (Bar-Ilan et al., 2009; Joachims et al., 2007), it has been shown that when the number of results is small than this influence is insignificant (Saracevic, 2007). Interestingly, Table 1 below shows that in all cases users changed their minds about the rankings. However, as also shown in Table 1, for three out of the four tasks there was no noticeable correspondence between the order of the results' presentation to the users and their ranks. Only for the Alzheimer Google&Bing task two of the first three displayed results were ranked at top-3 ranks by the users. This can be explained by the fact that in comparison to the other three tasks, for this task there were substantially less non-relevant results, with relevance value of 1 (by 53-123% in the first round of the experiment, on average over all the users, and by 15%-113% on average for the second round). Also, for the Alzheimer Google&Bing task the percentage of relevant results (with the relevance grade of 4) is substantially higher than for the other tasks (by 10-13%, on average for the first round, and by 15-24%, on average for the second round). The differences between the percentages of the most relevant and the least relevant results for every task and round are presented in Figures 1 and 2.

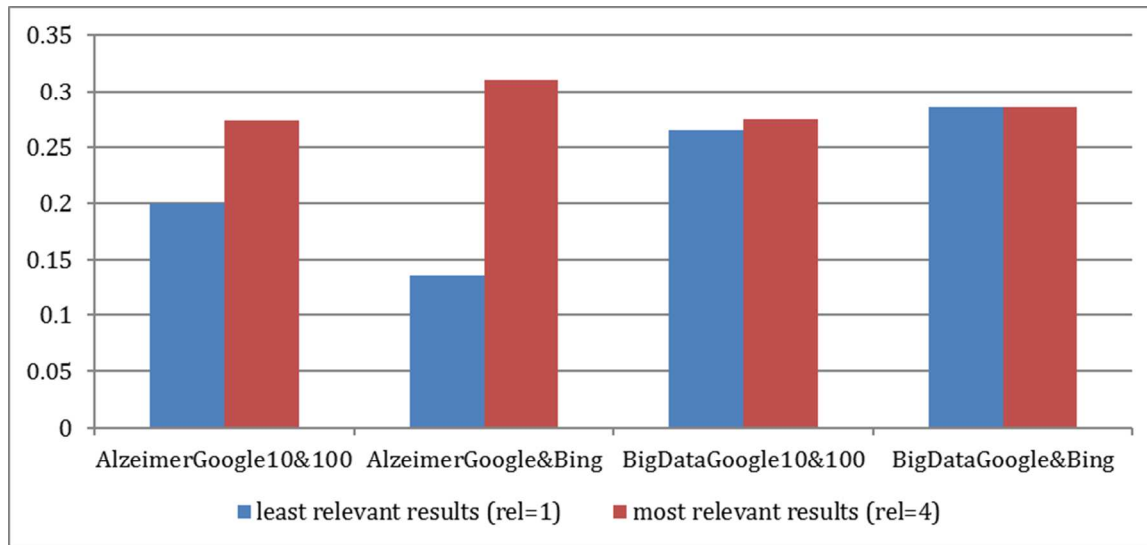


Figure 1: The averaged (on all the users) percentage of the results judged as least and most relevant for every task for the first round of the experiment.

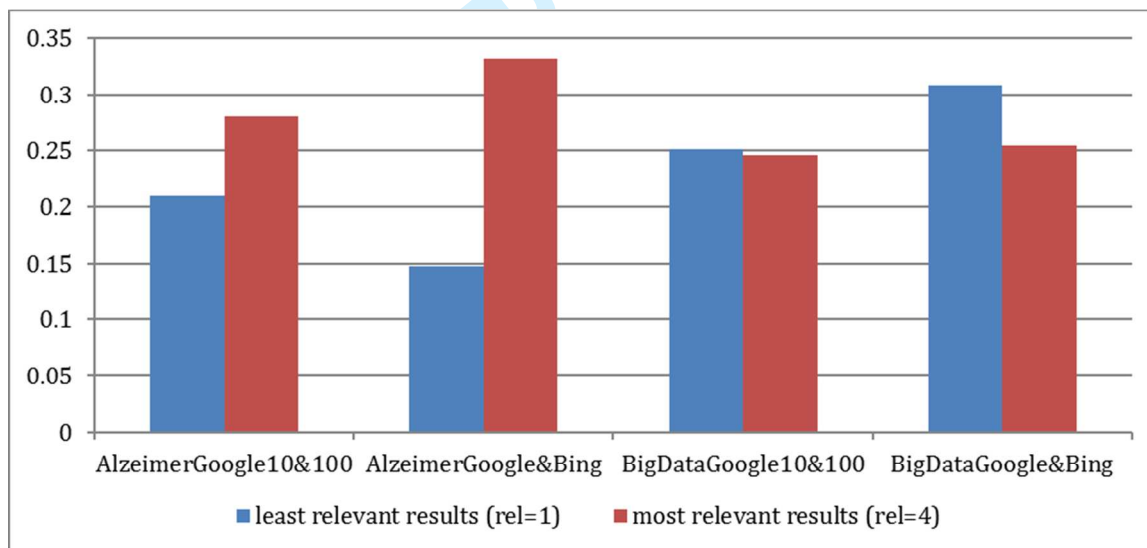


Figure 2: The averaged (on all the users) percentage of the results judged as least and most relevant for every task for the second round of the experiment.

Table 1. The average ranks of the first three results displayed to the users at each of the rounds for every query. The corresponding search engines' ranks of these results are displayed in parentheses (Google rank, Bing rank where available), if not available it is denoted as n/a.

	BigData Google&Bing	BigData Google10&100	Alzheimer Google&Bing	Alzheimer Google10&100
--	------------------------	-------------------------	--------------------------	---------------------------

Displayed to users as number	Ranked in Round1 as	Ranked in Round2 as	Ranked in Round1 as	Ranked in Round2 as	Ranked in Round1 as	Ranked in Round2 as	Ranked in Round1 as	Ranked in Round2 as
1	13 (5,n/a)	5 (4,n/a)	20 (6)	6 (3)	3 (7,7)	2 (4,3)	8 (101)	20 (106)
2	3 (8,n/a)	11 (7,n/a)	14 (105)	8 (110)	2 (n/a,10)	6(n/a,12)	2 (6)	14 (102)
3	7 (6,n/a)	14 (5,n/a)	16 (106)	9 (10)	15 (10,n/a)	3 (n/a,8)	14 (109)	11 (105)

The participants were informed that their rankings will be aggregated and analyzed anonymously, and those who wished not to contribute their data to the aggregated study were asked to inform the course instructor by email. No students asked to withdraw their data. Although the experiments involve human subjects (students), no personal information was gathered on them. The Faculty of Humanities’ IRB (ethics committee) waived the need for written consent. The IRB of the Faculty of Humanities in Bar-Ilan University approved the experiments.

3.3 Measures of change in user evaluation of search results

To test these research questions, two measures were proposed to calculate the aggregated judgments of search results that reflect the "wisdom of crowds" of a user group and compare their stability to the individual user judgments. Two types of judgments were considered in this study: 1) relevance on a 4-point scale with possible ties, and 2) ranking on a 10-point scale without ties.

To compute the aggregated “wisdom of crowds” ranking and relevance-based ranking grades, all the individual users' values for every result were summed up and the result list was sorted by these sums in ascending order to obtain the ranked list of results by users' ranking, and in descending order to obtain the list of results by users' relevance judgments. This was repeated for both rounds of the experiment. These aggregated results are referred to as *consensus ranking-based ranking* and *consensus relevance-based ranking*. In this study there were no ties (i.e. two items with exactly the same aggregate score). In case there are ties, these are resolved randomly.

To assess the stability of the judgments over time for each individual user and of the user consensus, two measures were devised. For each query and result set, the proportion of the results in the set that was *not* given identical ranks or relevance judgments by a specified user or by the user consensus, on the first and second rounds of the study was calculated. This measures the amount of change at the *exact match* level (i.e. results with distance 0 are those that were identically judged by the given user in both rounds).

Further, cases when the rankings or relevance judgments were not precisely identical in both rounds but still sufficiently close were also considered.

Formally, the *change coefficient* at a distance d , with $0 \leq d \leq |S|$ is defined for a given set of results, s_1, s_2, \dots, s_k , evaluated twice by a user (or by the user consensus), u , either with ranking or relevance values, r_1 and r_2 , as follows:

$$\Omega(d) = 1 - \frac{\sum_{i=1}^{|S|} n(s_i)}{|S|}, \text{ where } n(s_i) = \begin{cases} 1, & |r_1(s_i) - r_2(s_i)| \leq d \\ 0, & \text{otherwise} \end{cases}$$

Thus, $\Omega(d)$ is the proportion of the results that are judged in the two rounds at distance greater than d in the set S . Note that for $d=0$ the change coefficient reduces to the exact match case, while $d>0$ defines the more general case. For relevance all the 20 results in S were judged by the users and thus all of them are considered in the calculation of the change coefficient. However, for ranking only the top-10 results were actually assessed by the users. Therefore, for ranking only, as there are more results than ranks, the unranked results are technically assigned the rank of 11. Only results with at least one of the ranks being lower than 11 are considered. This is because results that were assigned rank 11 were not actually ranked by the users.

In addition, based on subsets of k ranks, the proportion of new *non-overlapping* results in the subset of k consecutive ranks is measured, which starts at a position p , that were introduced in the second round of the study. More formally, given a set of results R , and a consecutive subset of ranks $(r_p, r_{p+1}, \dots, r_{p+k})$ where $1 < r_i \leq |R|$, two subsets of ranked results are constructed with the corresponding ranks for each of the two rounds, R_1 and R_2 . Thus, the *change in k -subset* measure, is defined as follows:

$$NO(p, k) = 1 - |R_1 \cap R_2|/k, \text{ for some } p, k = \{1..N\}$$

In the sequel $NO(\text{top-}k)$ will stand for $NO(1, k)$ and $NO(\text{last-}k)$ will stand for $NO(N-k+1, k)$. This measure computes the proportion of results in a certain subset in one of the rounds that were not part of this subset in the other round.

4. Results and discussion

To answer the first research question of this study we computed the changes in the individual users' judgments and the corresponding changes in the consensus judgments by the two types of measures defined above and then compared these changes. It was found that the majority of changes in user evaluation of search results are local within 1-2 close ranks and relevance values. Moreover, the consensus rankings (obtained by aggregation of the individual ranks) were considerably more stable (changed in time less) than the individual users' rankings, but were still quite different (by 30-60%) from the search engine's rankings. These findings imply that "wisdom of crowds" decreases the subjectivity

and increases the stability of user ranking and thus can be used as a reliable reference for user relevance evaluation behaviour modelling.

4.1 Analysis of the change coefficient for individual user judgments vs. consensus judgments

To this end, first, the changes of individual users' rankings and relevance judgments for the same 20 results between the first and second rounds of the experiment were computed. To this end, the average of the change coefficients, $\Omega(d)$, with $d=0$ over the individual users' rankings and relevance judgments were calculated. The results for different studies are presented in Table 2. Then, the change coefficient was measured for distances greater than or equal to one between judgments in the two rounds. In our experiment $d=1$ for relevance, and $1 \leq d \leq 3$ for rankings were used (as the distance between judgments of the results in the two rounds), and are also presented in Table 2. For $\Omega(d > 1)$ only changes of distance 2 or more were counted (i.e. if for example, an item was ranked 7th in the first round, and 9th or above, or 5th or below in the second round, then we consider it as a change). As mentioned above, all unranked items by the user received a virtual rank of 11.

Table 2. The average change coefficient values of individual users for ranking and relevance with different distances and result sets. Standard deviation values are shown in parentheses following the average.

Experiment	Ranking				Relevance	
	$\Omega(0)$	$\Omega(1)$	$\Omega(2)$	$\Omega(3)$	$\Omega(0)$	$\Omega(1)$
AlzheimerGoogle10&100	0.87(0.18)	0.61(0.23)	0.45(0.22)	0.33(0.19)	0.48(0.11)	0.12(0.07)
AlzheimerGoogle&Bing	0.87(0.18)	0.67(0.21)	0.50(0.20)	0.39(0.17)	0.53(0.09)	0.15(0.05)
BigDataGoogle10&100	0.84(0.18)	0.62(0.22)	0.44(0.20)	0.31(0.17)	0.52(0.14)	0.13(0.06)
BigDataGoogle&Bing	0.87(0.18)	0.64(0.21)	0.48(0.20)	0.32(0.17)	0.43(0.15)	0.10(0.06)
Average of the averages	0.86(0.18)	0.64(0.22)	0.47(0.20)	0.34(0.17)	0.49(0.12)	0.13(0.06)

It can be observed that, in general, similarly high values (84-87% for ranking and 43-53% for relevance) were obtained for the different queries and result sets, recalling that when $\Omega(d)=0$ then no change occurred. Further, the explored research question is whether consensus ranking, which aggregates all the individual ranks for a given result into a single score, would exhibit a smaller amount of changes than the average for individual users. To test this question, each result for every query is assigned an identifying number. Then, the consensus rank/relevance score for every result of each query and result set is computed as a sum of all its individual user ranks/relevance grades, similar to the "agreed" ranking defined by Bar-Ilan et al. (2007). Then, all the results with a consensus rank higher than 10 were assigned a rank value of 11 as was done for the individual rankings, since in our experiments users could only rank the best 10 out 20 results. According to the definition of $\Omega(0)$ and similarly to the analysis of the

individual user judgments the change coefficients for the consensus ranking is calculated over the top-10 results only. The same method was applied to compute the consensus relevance grades for relevance-based ranking but all 20 results were assigned a rank in this case. Thus, the consensus relevance-based ranking shows a way to create a ranking for the full result list without the users ranking them explicitly. The results of the change coefficient measure for the consensus ranking and relevance-based ranking are displayed in Table 3.

The change coefficient for relevancies is considerably lower than for rankings (in Table 2), which may indicate that for users, ranking is generally more difficult than judging relevance. As expected, a consistent decrease in the change coefficient is observed, especially the considerable decrease between distance 0 and distance 1, for the relevance judgments, which indicates that most of the changes in relevance judgments were local within distance 1. Also, for ranking the majority of changes were local within distance 2. In addition, virtually similar numbers for both queries and result sets were obtained. This reflects a general pattern in user evaluation behaviour which is not specific to a specific case, users or data set.

Here we introduced new measures and a unique experimental setting, aiming to examine the time as the only varying parameter. The only closely related works for comparison are those by Scholer et al. (2011) and Scholer et al. (2013), which we mentioned earlier in the related work section. Our results on change in relevance with $d=0$ seem quite similar to those of (Scholer et al., 2013) who reported about 50% change rate. When considering $d=1$ as an approximation of the two-point scale, our results (12-15% change coefficients) are also comparable but slightly lower to those of (Scholer et al., 2011), who reported on 15-24% change rates. However, as opposed to our approach, Scholer et al. (2011) viewed cases where the document was evaluated differently the second time as errors. In the experimental setting of (Scholer et al., 2013), three documents were assessed twice within a short period of time (of about one hour) and the results in this study were not interpreted.

Table 3. The change coefficient for consensus ranking and relevance grades with various distances for different experiments.

Experiment	Ranking-based ranking				Relevance-based ranking			
	$\Omega(0)$	$\Omega(1)$	$\Omega(2)$	$\Omega(3)$	$\Omega(0)$	$\Omega(1)$	$\Omega(2)$	$\Omega(3)$
AlzheimerGoogle10&100	0.70	0.40	0.20	0.10	0.70	0.40	0.20	0.10
AlzheimerGoogle&Bing	0.80	0.60	0.40	0.40	0.85	0.50	0.50	0.40
BigDataGoogle10&100	0.60	0.30	0.10	0.10	0.70	0.40	0.20	0.10
BigDataGoogle&Bing	0.60	0.10	0.10	0.10	0.65	0.25	0.00	0.00
Average	0.68 (0.10)	0.35 (0.21)	0.20 (0.15)	0.18 (0.15)	0.73 (0.09)	0.39 (0.11)	0.23 (0.21)	0.15 (0.18)

The results for consensus ranking and relevance-based ranking are quite similar despite the fact that the former was calculated for top-10 results only, while the latter considers all 20 results. For ranking with $d=0$ (exact match) were obtained, while for $d>0$ the numbers are strictly monotonically decreasing (except for AlzheimerGoogle&Bing task for $\Omega(2)$ and $\Omega(3)$). As can be observed by comparing the first four columns of Tables 2 and 3, the numbers in Table 3 are 22% lower on average, for the averaged individual ranking change coefficient for the exact match, and 51% lower on average for $d>1$, than the corresponding values, shown in Table 2. The smallest decrease in change (of 8%) was observed for AlzheimerGoogle&Bing task ($d=0$), and the greatest decrease (of 84%) was observed for the consensus ranking of the BigDataGoogle&Bing task ($d=1$). The change coefficient for consensus ranking is more than one standard deviation lower than the average mean change coefficient for the individual users.

4.2 Analysis of the change in the subsets of k ranks for individual user judgments vs. consensus judgments

Next, our second measure was applied to compute the change in the subsets of k ranks. Table 4 considers the change in the top-5, last-5 and unranked results. Thus, the non-local (inter-subset) changes were measured between the two evaluation rounds in the top-5 result subset, (i.e. $NO(top-5)$), and in the last-5 result subset, (i.e. $NO(last-5)$), and in all the unranked results, (i.e. $NO(unranked)$), results that were unranked at least in one of the rounds, as a third subset of ranks.

Table 4. The change in ranking for unranked, top-5 and last 5 for the different experiments. Standard deviation values are shown in parentheses following the average.

	<i>NO(unranked)</i>	<i>NO(last-5)</i>	<i>NO(top-5)</i>
AlzheimerGoogle10&100	0.30(0.16)	0.64(0.28)	0.41(0.20)
AlzheimerGoogle&Bing	0.31(0.15)	0.61(0.26)	0.47(0.24)
BigDataGoogle10&100	0.25(0.14)	0.62(0.24)	0.39(0.26)
BigDataGoogle&Bing	0.23(0.14)	0.53(0.22)	0.41(0.22)
Average	0.27(0.15)	0.60(0.25)	0.42(0.24)

We note that for all the tasks there is much more change in the middle category subset of results (last-5) than in either the top ranked most relevant category (top-5) or the unranked least relevant subset. A reasonable explanation for this is that it is probably easier for the users to judge the extremes than the middle-category results. The most stable category was “unranked results” for which over two thirds of the

results remained unranked in both rounds. For the top-5 results the majority remained in top-5 over both rounds.

Table 5. The change in top-5 and last-5 and unranked for consensus ranking-based ranking for the different experiments.

	<i>NO(unranked)</i>	<i>NO(last-5)</i>	<i>NO(top-5)</i>
AlzheimerGoogle10&100	0.10	0.20	0.00
AlzheimerGoogle&Bing	0.10	0.60	0.40
BigDataGoogle10&100	0.10	0.20	0.00
BigDataGoogle&Bing	0.10	0.20	0.00
Average	0.10(0.0)	0.30(0.20)	0.10(0.20)

The $NO(p,k)$ values for the consensus ranking-based ranking also decreased in comparison to the individual user ranking as shown in Table 5 and compared to Table 4. The obtained results show less change in the top- k values for the top-5 consensus ranking (10% average $NO(top-5)$) than for the last-5 (30% average $NO(last-5)$). Moreover, there were lower changes except for the AlzheimerGoogle&Bing task in the top-5 and last-5 categories for the consensus ranking than for the individual user rankings in the top-5 results (where 42% of results were new in the second round on average for all the experiments). In particular, it was found that except for the Alzheimer Google&Bing task, there were no new results within the top-5 results in the consensus rankings between the rounds; however, there are some changes in the actual rankings. Similarly, only one out of the top-10 results (10%) of the consensus ranking was different for the first and second rounds for all the experiments (compared to 27% on average for individual users' ranking in Table 4). Again, as for the change coefficient above, the change in subsets of k ranks for consensus ranking is more than one standard deviation lower than the average mean change for the individual users.

Thus, the consensus ranking which reflects the “wisdom of crowds” evaluation is more stable than individual user rankings. Within distance 2 about 80% of consensus ranks on average did not change, and for top-5 virtually all the results remained in the top-5 subset in both rounds. The highest proportion of non-local changes and its lower decrease for consensus ranking observed for Alzheimer Google&Bing task may be explained by the highest proportion of relevant results which made the ranking task more difficult for the users for this query as discussed above; see Figures 1 and 2.

4.3 User consensus ranking versus search engine ranking

To address the second research question of this study we compared the user consensus ranking to the search engine rankings.

As was shown in the previous section, consensus ranking-based ranking is much more stable and less subjective than individual user-based ranking. Thus, it could serve as a good reference/gold standard for evaluating search engines' rankings. To this end we compared the user consensus rankings to the search engine rankings by using the change in k -subset measure to assess the difference between these two types of ranking. The results are shown in Table 6.

Table 6. The change in top- k for $k=10$ (all the ranked results) and for $k=5$ between the consensus ranking and the search engine ranking.

	Google- NO (top-10) round1, round2	Google- NO (top-5) round1, round2	Bing - NO (top-10) round1, round2	Bing - NO (top-5) round1, round2
AlzheimerGoogle10&100	0.20, 0.30	0.60, 0.60	N/A	N/A
AlzheimerGoogle&Bing	0.50, 0.50	0.60, 0.60	0.30, 0.30	0.60, 0.60
BigDataGoogle10&100	0.40, 0.30	0.40, 0.40	N/A	N/A
BigDataGoogle&Bing	0.30, 0.40	0.40, 0.40	0.50, 0.60	0.60, 0.60

For every experiment in the corresponding cell of the table the proportion of non-overlapping results is shown in the subset of top-10 and of top-5 computed for each of the two rounds separated by comma.

In general, from Table 6 we can see that the difference between users' consensus and search engines' rankings in both rounds is quite considerable (20-60% for the top-10 ranks) as has also been shown in a previous study (Bar-Ilan et al., 2007). Figures 3-6 show the overlap and the changes in the consensus rankings between rounds, with information added regarding the rankings assigned by the search engines. The overlapping results are inter-linked with arrows, while results that were ranked in the top-10 only for one of the rounds are marked with an X icon. Moreover, the difference between users' consensus and search engines' ranking is much higher than the change between the users' consensus rankings in the two rounds of the experiment (10%), as can be seen from Figures 3-6. However, the majority of results ranked in top-10 by Google were also ranked in top-10 by user consensus (again with exception for AlzheimerGoogle&Bing result set). Interestingly, comparable numbers were obtained for Google and Bing, while for one of the queries Google's ranking was closer to the consensus ranking than Bing's one (0.30 vs. 0.50, respectively), and for the other query Bing's ranking was closer to the consensus ranking (0.30 vs. 0.50, respectively) than Google's one.

Round 1		Round 2
1. Hebrew Wikipedia (Google 1)	→	1. Hebrew Wikipedia (Google 1)
2. Israeli Neurology Portal (Google 6)	↗	2. National Institute for Neurological Disorders in English (Google 103)
3. Alzheimer's Association site in English (Google 8)	→	3. Alzheimer's Association site in English (Google 8)
4. National Institute for Neurological Disorders in English (Google 103)	↘	4. Infomed - Israeli Medical Portal (Google 4)
5. Infomed - Israeli Medical Portal (Google 4)	↗	5. Israeli Neurology portal (Google 6)
6. Israeli Alzheimer's Association (Google 3)	↘	6. Living with Alzheimer's (Google 7)
7. Living with Alzheimer's (Google 7)	↗	7. Israeli Alzheimer's Association (Google 3)
8. Coping with Alzheimer's (Google 101)	✗	8. Health services in Israel: Alzheimer's (Google 107)
9. Info center for the elderly: Alzheimer's (Google 5)	→	9. Info center for the elderly: Alzheimer's (Google 5)
10. Voices of the soul: Alzheimer's (Google 105)	✗	10. Coping with Alzheimer's (Google 101)

Figure 3. Top-10 results of the consensus rankings for the Alzheimer Google 10&100 experiment. The corresponding Google ranks are shown in the parentheses. The change coefficient of users' consensus ranking-based ranking in two assessment rounds for this task is 0.70.

Round 1		Round 2
1.Hebrew Wikipedia (G 1; B1)	→	1.Hebrew Wikipedia (G 1; B 1)
2.Nursing homes network (G: N/A; B 10)	→	2.Infomed:The Israeli Medicine Portal (G 4; B 3)
3.Portals for Living with Alzheimer (G 7; B 7)	→	3.The Hebrew Disease Index: Alzheimer's (G N/A; B 8)
4.Alzheimer's Association site in English (G 8; B N/A)	→	4.Israeli Neurology Portal (G 6; B N/A)
5.Infomed:The Israeli Medicine Portal (G 4; B 3)	→	5.Alzheimer Association site in English(G 8; B N/A)
6. The Hebrew Disease Index: Alzheimer's (G N/A; B 8)	→	6. Clalit Health Services – Alzheimer's(G N/A; B 12)
7. Take care – Alzheimer (G N/A; B 5)	→	7. Take care – Alzheimer (G N/A; B 5)
8. Alzheimer's portal (G N/A; B 2)	→	8. Portal for Living with Alzheimer (G 7; B 7)
9. RamatGan Center for Alzheimer (G13;BN/A)	→	9. Alzheimer's portal (G N/A; B 2)
10. Israeli Neurology Portal (G 6; B N/A)	→	10. Nursing homes network (G: N/A; B 10)

Figure 4. Top-10 results of the consensus rankings for the Alzheimer G&B experiment. The corresponding Google (G) and Bing (B) ranks are shown in the parentheses. The change coefficient of users' consensus ranking-based ranking in two assessment rounds for this task is 0.80.

Round 1		Round 2
1. English Wikipedia (Google 1)	→	1.English Wikipedia (Google 1)
2. YouTube video on Big Data in English (Google 8)	→	2.YouTube video on Big Data in English (Google 8)
3. SAS page in English (Google 4)	→	3.Hebrew Wikipedia (Google 2)
4. Hebrew Wikipedia (Google 2)	→	4.Webopedia in English (Google 9)
5. Webopedia in English (Google 9)	→	5.SAS page in English (Google 4)
6. McKinsey – Big data (Google 3)	→	6. McKinsey – Big data (Google 3)
7. NITRD group – big data (Google 109)	→	7. NITRD group – big data (Google 109)
8. Intel – Big data analytics(Google 107)	→	8.World Economic Forum(Google 110)
9. Qualitest blog (Google 104)	→	9.IBM Big data platform (Google 10)
10.World Economic Forum (Google 110)	→	10.Qualitest Blog (Google 104)

Figure 5. Top-10 results of the consensus ranking for the Big Data Google 10&100 task. The corresponding Google ranks are shown in the parentheses. The change coefficient of users' consensus ranking-based ranking in two assessment rounds for this task is 0.60.

Round 1		Round 2
1.O'REILLY's page in English (G N/A; B 7)		1. English Wikipedia (G1; B1)
2. English Wikipedia (G 1; B 1)		2.O'REILLY's page in English(G N/A; B 7)
3. YouTube video on Big Data in English (G 8; B N/A)		3.YouTube video on Big Data in English (G 8; B N/A)
4. SAS page in English (G 4; B N/A)		4.Hebrew Wikipedia (G2; B2)
5. Hebrew Wikipedia (G 2; B 2)		5.SAS page in English (G 4; B N/A)
6. What is big data? (G N/A; B 9)		6. What is big data? (G N/A; B 9)
7. McKinsey – Big data (G 3; B N/A)		7. McKinsey – Big data (G 3; B N/A)
8. Webopedia in English (G 9; B N/A)		8. Webopedia in English (G 9; B N/A)
9. IBM - What is big data (G N/A; B 12)		9. A day in big data (G 11; B N/A)
10. A day in big data (G 11; B N/A)		10. Big data, big decisions (G N/A; B 6)

Figure 6. Top-10 results of the consensus ranking for the Big Data G&B experiment. The corresponding Google (G) and Bing (B) ranks are shown in the parentheses. The change coefficient of users' consensus ranking-based ranking in two assessment rounds for this task is 0.60.

Interestingly, for all the queries, 2 to 3 results were overlapping in the top-5 ranks for Google, Bing, and the user consensus ranking on both rounds. Also, all the results that were not ranked in the top-10 in the second round appeared at low ranks (rank 8 or lower) in the first round, which reflects higher stability in the top-7 consensus rankings than in the lower ones. For three out of four tasks in both rounds, the top result was the same as that of the search engines and for the user consensus rankings. Only for one task in the first round, was the search engine's top-ranked item ranked as second. As can be observed from the Figures 3-6, there were smaller differences in the top-10 consensus ranks for the Big Data query than for the Alzheimer's query, especially for the Alzheimer Google&Bing task, which appeared to be the most controversial and least stable according to the applied measures.

Users generally preferred English sites, when available, over Hebrew ones, and in one instance a YouTube video outranked textual results. For Bing there were cases of irrelevant results referring to a different meaning of Big Data (like the music project and a recruitment management company). Quite surprisingly, for the Alzheimer Google10&100 task three out of top-10 results, and for Big Data Google10&100 four out of top-10 results in the consensus rankings were from Google's 100+ results set. These findings show that even on the tenth SERP there may be results that may be preferred to those in the top-10 shown to the users.

5. Conclusions and future work

Ranking of search results according to their relevance to the users is one of the primary tasks of search engines. However, this task is extremely challenging especially due to the changes over time in user preferences, which affect their assessment of search results. This paper presented an exploratory study of this issue that has not been addressed before.

The primary goal of this study was to investigate whether and how the user preferences change over time. In particular, the main idea was to test whether aggregated consensus judgments are more stable than individual user judgments and whether they are similar to the search engines' rankings. To this end, two new measures of change were proposed for ranking and relevance judgments, the change coefficient, $\Omega(d)$, at distance d , and the change in k -subset measure $NO(p,k)$, for a consecutive subset k of ranks.

To aid our investigation, a large-scale user study was conducted for ranking and judging the relevance of query result sets, and repeated the evaluation within a two-month period. It was found that the amount of changes was quite high for individual users, but the majority of changes both for relevance and ranking judgments were local within distance 1 and 2, respectively. In addition, the overlap in the top-5 and unranked subsets of the results was higher than in the last-5 subset, which implies that users are more certain about ranking of the top and least relevant results than ranking of the middle subset.

In addition, our results show that consensus ranking calculated by aggregating the individual user rankings and relevance judgments resulted in substantially fewer changes compared to the averaged individual user rankings. Finally, as in other studies (Bar-Ilan and Levene, 2007), quite low similarity was found between the search engines' and consensus users' rankings.

Generally, understanding the user intent is beneficial, as in some cases the most relevant results may be further down the ranking list (as shown by our Google 10&100 tasks). The above tendencies were quite similar for all the queries and result sets.

This is a user study with a relatively large number of participants, however, the main limitation of this study is the fact that the participants only judged the results of two informational queries they were asked to assess. In addition, this study is based on a user behaviour model, where users evaluate only a limited amount of results (20) coming from one or two search engines. The conclusions could be generalised by experimenting with more queries from a broad spectrum of topics, and also in the context of library and information science, where information seeking has been widely researched (Case, 2012).

The findings of this research contribute to understanding the user evaluation behaviour and its change over time and show a way to bridge the gap between search engines' and users' ranking and relevance evaluation. This research may have practical implications for personalisation, as users' preferences change over time and therefore the ranking of a search engine should adapt to this. For ranking, with the maximal locality threshold ($d=3$), about one third of the results undergo non-local changes, so it is these differently evaluated results that are especially in need of personalisation.

5.1 Implications of the study

Personalisation of search engine results has been much researched in the past decade (Keenoy and Levene, 2005; Micarelli et al., 2007), although it is unclear what the uptake of personalisation has been in commercial search services, mainly due to the scale of the problem and the unclear benefits of such an undertaking. It is well-known that the automated retrieval algorithms used by search engines take into account the popularity of user choices from analysis of the click-through data it records, however, the details of how this is done remain undisclosed. The results in this paper may provide insight to several aspects related to personalisation of search. As we have shown, there is a considerable difference between search engines' ranking and users' consensus ranking. From this we may conclude that search engines ranking of results is not fully compatible with users' preferences. Moreover, we have shown that users' judgements change considerably in time (over 30% of the results have non-local changes), so this change in users' preferences is another important factor that should be taken into account when ranking search results. Personalisation on an individual level may be viable in e-commerce and internet advertising, where the benefits are clearly visible, although for search engines there is no proven model for personalisation as yet. Nonetheless, for search engines, taking the consensus ranking into account may be a reasonable solution to improving the quality of results' ranking. What we have clearly shown in the paper is that the consensus ranking is more stable than individual rankings, so apart from the computational benefits of such an approach, it is more stable than dealing with individual users where the changes are more variable and thus less predictable. There is also another positive side to the consensus ranking in that it will most likely result in a more diverse search results list (Santos et al., 2015), than would arise from personalisation on an individual level.

References

Agichtein, E., E. Brill and S. Dumais. (2006). "Improving Web search ranking by incorporating user behavior information". In *Proceedings of SIGIR'06*, pp. 19-26, ACM: New York, NY, USA.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Bao, S. G., Xue, X. Wu, Y. Yu, B. Fei and Z. Su. (2007). "Optimizing Web search using social annotations". In P. Patel-Schnider, P. Shenoy, C. Williamson, & M. Zurko (Eds.), *In WWW '07: Proceedings of the 16th International Conference on World Wide Web*, pp. 501-510, New York, NY.

Bates, M. (1989). "The design of browsing and berrypicking techniques for the online search interface". *Online Review*, Vol. 13 No. 5, pp. 407-424.

Bar-Ilan J, Keenoy K, Yaari E, Levene M. (2007). "User rankings of search engine results". *Journal of the Association for Information Science and Technology*, Vol. 58 No. 9, pp. 1254-1266.

Bar-Ilan J, Keenoy K, Yaari E, Levene M. (2009). "Presentation bias is significant in determining user preference for search results – A user study". *Journal of the American Society for Information Science and Technology*, Vol. 60 No. 1, pp. 135-149.

Bar-Ilan J, Levene M. (2011). "A method to assess search engine results". *Online Information Review*, Vol. 35 No. 6, pp. 854-868.

Bateman, J. (1998). "Changes in relevance criteria: A longitudinal study". *Journal of the American Society for Information Science*, Vol. 35, pp. 23–32.

Bilal, D. (2012), "Ranking, relevance judgment, and precision of information retrieval on children's queries: Evaluation of Google, Yahoo!, Bing, Yahoo! Kids, and ask Kids". *Journal of Association for Information Science*, 63: 1879–1896. doi: 10.1002/asi.22675.

Bollen, J. and H. Mao. (2011). "Twitter Mood as a Stock Market Predictor". *IEEE Computer*, Vol. 44 No. 10, pp. 91-94.

Bruce, H.W. (1994). "A cognitive view of the situational dynamism of user centered relevance estimation". *Journal of the Association for Information Science*, Vol. 45 No. 5, pp. 142–148.

Case, D. O. (2012). *Looking for information: A survey of research on information seeking, needs and behavior*. Bingley, UK, Emerald Group Publishing Limited.

Cen, R., Y. Liu, M. Zhang, L. Ru and S. Ma. (2009). "Automatic search engine performance evaluation with the wisdom of crowds". In *Proceedings of AIRS 2009*, Japan.

Chitika (2013). Chitika insights: The value of Google positioning. Retrieved from <https://cdn2.hubspot.net/hub/239330/file-61331237-pdf/ChitikaInsights-ValueofGoogleResultsPositioning.pdf>

Choochaiwattana, W. and Spring, M. B. (2009). "Applying social annotations to retrieve and re-rank Web resources". In Mahadevan V., Yi Xie (Eds.), In *ICIME '09: Proceeding of 2009 International Conference On Information Management and Engineering*, pp. 215-219, Kuala Lumpur, Malaysia, 3-5 April 2009. IEEE, Los-Alamitos, CA.

Cooper, S., Khatib, F., Treuille, A., et al. (2010). "Predicting protein structures with a multiplayer online game". *Nature*, Vol. 466, pp. 756–60.

comScore (2015). comScore Releases November 2015 U.S. Desktop Search Engine Rankings.

Dervin, B. (1992). In: *Qualitative Research in Information Management*. Englewood, CO:Libraries Unlimited, pp. 61-84.

Dou Z., R. Song, X. Yuan, J. Wen. (2008). "Are click-through data adequate for learning web search rankings?" In *Proceedings of CIKM'08*, pp. 73- 82. ACM: New York, NY, USA.

Du, J.T. (2010). "Multitasking, cognitive coordination and cognitive shifts during Web searching". Unpublished Ph.D., Queensland University of Technology, Australia -- Queensland.

Du, J.T. & Spink, A. (2011). "Towards a Web search model: Integrating multitasking, cognitive coordination and cognitive shifts". *Journal of the American Society for Information Science and Technology (JASIST)*, 62(8), 1446–1472.

Ellis, D. (1989). "A behavioural approach to information retrieval design". *Journal of Documentation*, Vol. 49 No. 4, pp. 171-212.

Fisher, K. E., Erdelez, S. and McKechnie, L.E.F (Eds). (2005). *Theories of information behavior*. ASIS&T Monograph Series. Medford, NJ: Information Today.

Giles, G. (2005). "Internet encyclopedia go head to head". *Nature*, Vol. 438, pp. 900-901.

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

Harris, C. G. (2014, April). "The beauty contest revisited: measuring consensus rankings of relevance using a game". In *Proceedings of the First International Workshop on Gamification for Information Retrieval* (pp. 17-21). ACM.

Jansen B.J., Spink A. (2006). "How are we searching the Web? A comparison of nine search engine transaction logs". *Information Processing and Management*, Vol. 42, pp. 248-263.

Harshavardhan, A., Gandhe, A., Ross, L., Ssu-Hsin, Y., and L. Benyuan. (2013), "Online Social Networks Flu Trend Tracker: A Novel Sensory Approach to Predict Flu Trends". *Biomedical Engineering Systems and Technologies, Communications in Computer and Information Science*, Vol. 357, pp. 353-368.

Joachims T, Granka L, Pan B, Hembrooke H, Radlinks F, and Gay G. (2007). "Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search". *ACM Transactions on Information Systems*, Vol. 25 No. 2, Article 7.

Kawase, R., Siehdn, P., Pereira Nunes, B., Herder, E., & Nejdl, W. (2014). "Exploiting the wisdom of the crowds for characterizing and connecting heterogeneous resources". In *Proceedings of HT'14*, September 1-4, 2014, Santiago, Chile.

Keenoy K. and M. Levene (2005). "Personalisation of web search", In: *Intelligent Techniques for Web Personalization (ITWP)*, Eds. S.S. Anand and B. Mobasher, *Lecture Notes in Computer Science (LNCS)*, pp. 201- 228, Springer-Verlag, Berlin.

Knight, S. A. and Spink, A. (2008). "Toward a web search information behavior model". In A. Spink and M. Zimmer (Eds.) *Web Search: Multidisciplinary Perspectives*. Berlin: Springer-Verlag.

Lewandowski, D. (2008). "The retrieval effectiveness of web search engines: Considering results descriptions". *Journal of Documentation*, 64(6), 915-937.

Lewandowski, D. (2015), "Evaluating the retrieval effectiveness of web search engines using a representative query sample". *Journal of the Association for Information Science and Technology*, 66: 1763-1775. doi: 10.1002/asi.23304.

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge, UK:Cambridge University Press.

Micarelli A., F. Gasparetti, F. Sciarrone, and S. Gauch. (2007). Personalized Search on the World Wide Web. In P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.): *The Adaptive Web*, LNCS 4321, pp. 195–230. Springer-Verlag: Berlin Heidelberg.

Mizzaro S. (1998). “How many relevances in information retrieval?” *Interacting with Computers*, Vol. 10 pp. 305-322.

Mortensen, J. M., Minty E. P., Januszuk, M., Sweeney, T. E., Rector, A. L, Noy, N. F., & Musen, M. A. (2015). “Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT”. *Journal of American Medical Information Association*, Vol. 22 No. 3, pp. 640-648.

Preis, T., Moat H. S., H. E. Stanley. (2013), "Quantifying Trading Behavior in Financial Markets Using Google Trends", *Scientific Reports*, Vol. 3 No. 1684.

Rees, A.M., and Schultz, D.G. (1967). *A field experimental approach to the study of relevance assessments in relation to document searching* (vols.1–2). Cleveland, OH: Western Reserve University, School of Library Science, Center for Documentation and Communication Research.

Saracevic T. (1996). “Relevance reconsidered”. (1996). In *CoLIS 2: Proceedings of the Second Conference on Conception of Library and Information Science: Integration in Perspectives*; October 13-16 1996, Copenhagen, Denmark: The Royal School of Librarianship, pp. 201-218.

Santos, R. L. T. C. MacDonald, I. Ounis (2015).” Search Result Diversification”. *Foundations and Trends in Information Retrieval*, Vol. 9, No. 1 (2015) 1–90. DOI: 10.1561/15000000040.

Saracevic T. (2007). “Relevance: a review of the literature and a framework for thinking on the notion in information science, Part III: Behaviour and Effects of Relevance”. *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 13, pp. 2126-2144.

Scholer F, Turpin A, and Sanderson M. (2011). “Quantifying test collection quality based on the consistency of relevance judgments”. In *SIGIR’11: Proceedings of the 34th international ACM SIGIR conference*; July 24-28 2011; Beijin, China. New York: ACM; 2011. pp. 1063-1072.

Scholer, F., Kelly, D., Wu, W. C., Lee, H. S., and Webber, W. (2013). “The effect of threshold priming and need for cognition on relevance calibration and assessment”. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 623-632. ACM.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Singh, V. K., Piryani, R., Uddin, A., & Pinto, D. (2013). "A content-based e-resource recommender system to augment ebook-based learning". In *Multi-disciplinary Trends in Artificial Intelligence*, pp. 257-268. Springer Berlin Heidelberg.

Spink, A. (1977). "Study of interactive feedback during mediated information retrieval". *Journal of the American Society for Information Science*, Vol. 48 No. 5, pp. 382-394.

Smithson, S. (1994). "Information retrieval evaluation in practice: A case study approach". *Information Processing and Management*, Vol. 30 No. 2, pp. 205-221.

Surowiecki, J. (2005). *The wisdom of crowds*. New York: Doubleday.

Vakkari, P. and Hakala, N. (2000). "Changes in relevance criteria and problem stages in task performance". *Journal of Documentation*, Vol. 56 No. 5, pp. 540-562.

Vakkari, P. (2001). "Changes in search tactics and relevance judgments when preparing a research proposal: A summary of findings of a longitudinal study". *Information Retrieval*, Vol. 4 No. 3, pp. 295-310.

Vaughan, L. (2004). "New measurements for search engine evaluation proposed and tested, *Information Processing & Management*", 40(4), 677-691.

Veronis, J. (2006). "A comparative study of six search engines". Retrieved April 15, 2006, from <http://www.up.univ-mrs.fr/veronis/pdf/2006-comparative-study.pdf>.

Wang, P., and White, M.D. (1995). "Document use during a research project: A longitudinal study". *Proceedings of American Society for Information Science*, Vol. 32, pp. 181-188.

Yanbe, Y. A. Jatowt, S. Nakamura, & K. Tanaka. (2007). "Can social bookmarking enhance search in the Web? " In R. Larson, E. Rasmussen, S. Sugimoto, and E. Toms (Eds.), In *JCDL '07 Proceedings of the 2007 Conference on Digital Libraries*, pp. 107-116, New York, NY.

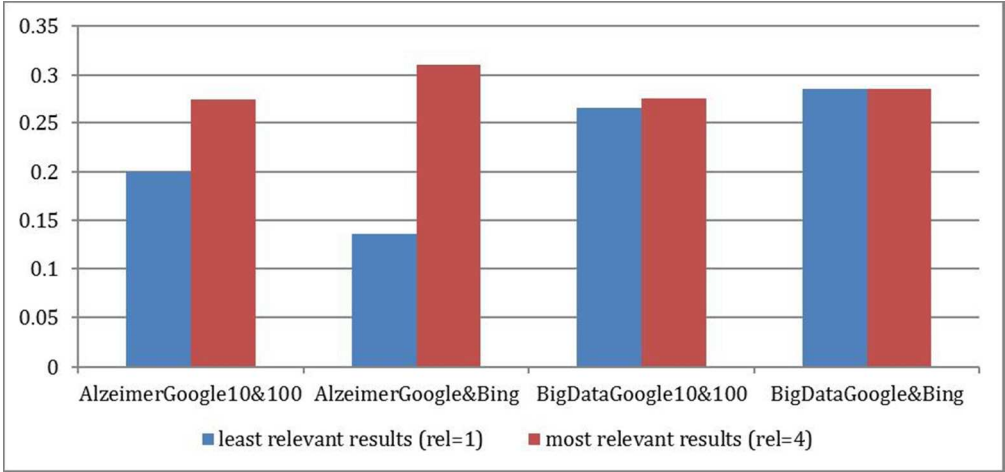
Zhang, Y. and Moffat, A. (2006). "Some observations on user search behavior". *Australian Journal of Intelligent Information Processing Systems*, Vol. 9 No. 2, pp. 1-8.

Zhitomirsky-Geffet M. and Y. Daya. (2015). "Mining query subtopics from social tags". *Information Research*, Vol. 20 No. 2, paper 66, retrieved from <http://InformationR.net/ir/20-2/paper666.html>.

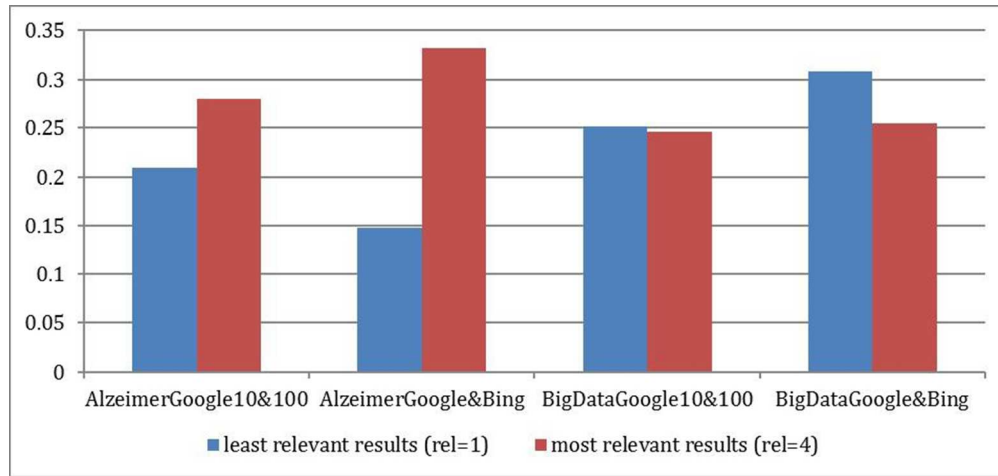
Zhitomirsky-Geffet, M., and Erez, E. S. (2014). "Maximizing agreement on diverse ontologies with "wisdom of crowds" relation classification". *Online Information Review*, Vol. 38 No. 5, pp. 616 - 633.

Zhitomirsky-Geffet M., Eden S. Erez, Judit Bar-Ilan. (2016). Towards Multi-viewpoint Ontology Construction by Collaboration of Non-experts and Crowdsourcing: The Case of the Effect of Diet on Health. *Journal of the Association for Information Science and Technology (JASIST)*. In press.

For Peer Review



The averaged (on all the users) percentage of the results judged as least and most relevant for every task for the first round of the experiment.
162x76mm (150 x 150 DPI)



The averaged (on all the users) percentage of the results judged as least and most relevant for every task for the second round of the experiment.
162x76mm (150 x 150 DPI)

Round 1		Round 2
1.Hebrew Wikipedia (Google 1)	→	1.Hebrew Wikipedia (Google 1)
2.Israeli Neurology Portal (Google 6)	↗	2.National Institute for Neurological Disorders in English (Google 103)
3.Alzheimer's Association site in English (Google 8)	↘	3.Alzheimer's Association site in English (Google 8)
4. National Institute for Neurological Disorders in English (Google 103)	↗	4.Infomed - Israeli Medical Portal (Google 4)
5.Infomed - Israeli Medical Portal (Google 4)	↘	5.Israeli Neurology portal (Google 6)
6. Israeli Alzheimer's Association (Google 3)	↗	6. Living with Alzheimer's (Google 7)
7. Living with Alzheimer's (Google 7)	↘	7. Israeli Alzheimer's Association (Google 3)
8. Coping with Alzheimer's(Google 101)	✗	8. Health services in Israel: Alzheimer's (Google 107)
9. Info center for the elderly: Alzheimer's (Google 5)	→	9. Info center for the elderly: Alzheimer's (Google 5)
10.Voices of the soul: Alzheimer's (Google 105)	✗	10. Coping with Alzheimer's (Google 101)

Top-10 results of the consensus rankings for the Alzheimer Google 10&100 experiment. The corresponding Google ranks are shown in the parentheses. The change coefficient of users' consensus ranking-based ranking in two assessment rounds for this task is 0.70.
171x127mm (96 x 96 DPI)











Round 1		Round 2
1. Hebrew Wikipedia (G 1; B1)	→	1. Hebrew Wikipedia (G 1; B 1)
2. Nursing homes network (G: N/A; B 10)	→	2. Infomed: The Israeli Medicine Portal (G 4; B 3)
3. Portal for Living with Alzheimer (G 7; B 7)	→	3. The Hebrew Disease Index: Alzheimer's (G N/A; B 8)
4. Alzheimer's Association site in English (G 8; B N/A)	→	4. Israeli Neurology Portal (G 6; B N/A)
5. Infomed: The Israeli Medicine Portal (G 4; B 3)	→	5. Alzheimer Association site in English (G 8; B N/A)
6. The Hebrew Disease Index: Alzheimer's (G N/A; B 8)	→	6. Clalit Health Services – Alzheimer's (G N/A; B 12)
7. Take care – Alzheimer (G N/A; B 5)	→	7. Take care – Alzheimer (G N/A; B 5)
8. Alzheimer's portal (G N/A; B 2)	→	8. Portal for Living with Alzheimer (G 7; B 7)
9. Ramat Gan Center for Alzheimer (G 13; B N/A)	→	9. Alzheimer's portal (G N/A; B 2)
10. Israeli Neurology Portal (G 6; B N/A)	→	10. Nursing homes network (G: N/A; B 10)

Top-10 results of the consensus rankings for the Alzheimer G&B experiment. The corresponding Google (G) and Bing (B) ranks are shown in the parentheses. The change coefficient of users' consensus ranking-based ranking in two assessment rounds for this task is 0.80.
173x111mm (96 x 96 DPI)

Round 1		Round 2
1. English Wikipedia (Google 1)	→	1. English Wikipedia (Google 1)
2. YouTube video on Big Data in English (Google 8)	→	2. YouTube video on Big Data in English (Google 8)
3. SAS page in English (Google 4)	↘	3. Hebrew Wikipedia (Google 2)
4. Hebrew Wikipedia (Google 2)	↗	4. Webopedia in English (Google 9)
5. Webopedia in English (Google 9)	↗	5. SAS page in English (Google 4)
6. McKinsey – Big data (Google 3)	→	6. McKinsey – Big data (Google 3)
7. NITRD group – big data (Google 109)	→	7. NITRD group – big data (Google 109)
8. Intel – Big data analytics (Google 107)	✗	8. World Economic Forum (Google 110)
9. Qualitest blog (Google 104)	↘	9. IBM Big data platform (Google 10)
10. World Economic Forum (Google 110)	↘	10. Qualitest Blog (Google 104)

Top-10 results of the consensus ranking for the Big Data Google 10&100 task. The corresponding Google ranks are shown in the parentheses. The change coefficient of users’ consensus ranking-based ranking in two assessment rounds for this task is 0.60.

170x86mm (96 x 96 DPI)

Round 1		Round 2
1.O'REILLY's page in English (G N/A; B 7)		1. English Wikipedia (G1; B1)
2. English Wikipedia (G 1; B 1)		2.O'REILLY's page in English(G N/A; B 7)
3. YouTube video on Big Data in English (G 8; B N/A)		3.YouTube video on Big Data in English (G 8; B N/A)
4. SAS page in English (G 4; B N/A)		4.Hebrew Wikipedia (G2; B2)
5. Hebrew Wikipedia (G 2; B 2)		5.SAS page in English (G 4; B N/A)
6. What is big data? (G N/A; B 9)		6. What is big data? (G N/A; B 9)
7. McKinsey – Big data (G 3; B N/A)		7. McKinsey – Big data (G 3; B N/A)
8. Webopedia in English (G 9; B N/A)		8. Webopedia in English (G 9; B N/A)
9. IBM - What is big data (G N/A; B 12)		9. A day in big data (G 11; B N/A)
10. A day in big data (G 11; B N/A)		10. Big data, big decisions (G N/A; B 6)

Top-10 results of the consensus ranking for the Big Data G&B experiment. The corresponding Google (G) and Bing (B) ranks are shown in the parentheses. The change coefficient of users' consensus ranking-based ranking in two assessment rounds for this task is 0.60.

172x85mm (96 x 96 DPI)

	BigData <i>Google&Bing</i>		BigData <i>Google10&100</i>		Alzheimer <i>Google&Bing</i>		Alzheimer <i>Google10&100</i>	
Displayed to users as number	Ranked in Round1 as	Ranked in Round2 as	Ranked in Round1 as	Ranked in Round2 as	Ranked in Round1 as	Ranked in Round2 as	Ranked in Round1 as	Ranked in Round2 as
1	13 (5,n/a)	5 (4,n/a)	20 (6)	6 (3)	3 (7,7)	2 (4,3)	8 (101)	20 (106)
2	3 (8,n/a)	11 (7,n/a)	14 (105)	8 (110)	2 (n/a,10)	6(n/a,12)	2 (6)	14 (102)
3	7 (6.n/a)	14 (5,n/a)	16 (106)	9 (10)	15 (10,n/a)	3 (n/a,8)	14 (109)	11 (105)

Experiment	Ranking				Relevance	
	$\Omega(0)$	$\Omega(1)$	$\Omega(2)$	$\Omega(3)$	$\Omega(0)$	$\Omega(1)$
AlzheimerGoogle10&100	0.87(0.18)	0.61(0.23)	0.45(0.22)	0.33(0.19)	0.48(0.11)	0.12(0.07)
AlzheimerGoogle&Bing	0.87(0.18)	0.67(0.21)	0.50(0.20)	0.39(0.17)	0.53(0.09)	0.15(0.05)
BigDataGoogle10&100	0.84(0.18)	0.62(0.22)	0.44(0.20)	0.31(0.17)	0.52(0.14)	0.13(0.06)
BigDataGoogle&Bing	0.87(0.18)	0.64(0.21)	0.48(0.20)	0.32(0.17)	0.43(0.15)	0.10(0.06)
Average of the averages	0.86(0.18)	0.64(0.22)	0.47(0.20)	0.34(0.17)	0.49(0.12)	0.13(0.06)

For Peer Review

Experiment	Ranking-based ranking				Relevance-based ranking			
	$\Omega(0)$	$\Omega(1)$	$\Omega(2)$	$\Omega(3)$	$\Omega(0)$	$\Omega(1)$	$\Omega(2)$	$\Omega(3)$
AlzheimerGoogle10&100	0.70	0.40	0.20	0.10	0.70	0.40	0.20	0.10
AlzheimerGoogle&Bing	0.80	0.60	0.40	0.40	0.85	0.50	0.50	0.40
BigDataGoogle10&100	0.60	0.30	0.10	0.10	0.70	0.40	0.20	0.10
BigDataGoogle&Bing	0.60	0.10	0.10	0.10	0.65	0.25	0.00	0.00
Average	0.68	0.35	0.20	0.18	0.73	0.39	0.23	0.15
	(0.10)	(0.21)	(0.15)	(0.15)	(0.09)	(0.11)	(0.21)	(0.18)

	<i>NO(unranked)</i>	<i>NO(last-5)</i>	<i>NO(top-5)</i>
AlzheimerGoogle10&100	0.30(0.16)	0.64(0.28)	0.41(0.20)
AlzheimerGoogle&Bing	0.31(0.15)	0.61(0.26)	0.47(0.24)
BigDataGoogle10&100	0.25(0.14)	0.62(0.24)	0.39(0.26)
BigDataGoogle&Bing	0.23(0.14)	0.53(0.22)	0.41(0.22)
Average	0.27(0.15)	0.60(0.25)	0.42(0.24)

For Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

	<i>NO(unranked)</i>	<i>NO(last-5)</i>	<i>NO(top-5)</i>
AlzheimerGoogle10&100	0.10	0.20	0.00
AlzheimerGoogle&Bing	0.10	0.60	0.40
BigDataGoogle10&100	0.10	0.20	0.00
BigDataGoogle&Bing	0.10	0.20	0.00
Average	0.10(0.0)	0.30(0.20)	0.10(0.20)

For Peer Review

	Google–NO(top-10) round1, round2	Google–NO(top-5) round1, round2	Bing – NO(top-10) round1, round2	Bing – NO(top-5) round1, round2
AlzheimerGoogle10&100	0.20, 0.30	0.60, 0.60	N/A	N/A
AlzheimerGoogle&Bing	0.50, 0.50	0.60, 0.60	0.30, 0.30	0.60, 0.60
BigDataGoogle10&100	0.40, 0.30	0.40, 0.40	N/A	N/A
BigDataGoogle&Bing	0.30, 0.40	0.40, 0.40	0.50, 0.60	0.60, 0.60

For Peer Review