Analysis of change in users' assessment of search results over time

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 of London
London, UK
Mark@dcs.bbk.ac.uk

*Corresponding author

E-mail: maayan.zhitomirsky-geffet@biu.ac.il

Abstract

We present the first systematic study of the influence of time on user judgements for rankings and relevance grades of web search engine results. The goal of this study is to evaluate the change in user assessment of search results and explore how users' judgements change. To this end, we conducted a large-scale user study with 86 participants who evaluated two different queries and four diverse result sets twice with an interval of two months. To analyse the results we investigate whether two types of patterns of user behaviour from the theory of categorical thinking hold for the case of evaluation of search results: (1) coarseness and (2) locality. To quantify these patterns we devised two new measures of change in user judgements and distinguish between local (when users swap between close ranks and relevance values) and non-local changes. Two types of judgements were considered in this study: 1) relevance on a 4-point scale, and 2) ranking on a 10-point scale without ties. We found that users tend to change their judgements of the results over time in about 50% of cases for relevance and in 85% of cases for ranking. However, the majority of these changes were local.

Introduction

Search engines provide users with a major gateway to the vast amount of information available on the Internet. For almost every query there are numerous results matching the query, yet users are usually content only with the 10-20 items displayed on the first results page (Jansen and Spink, 2006). The top search engine results are presumed to be the most relevant for the given query. However, previous research reveals quite a high level of disagreement between the ranking of search engines and user-produced rankings (Bar-Ilan et al., 2007; Bar-Ilan and Levene, 2011). Inter-user agreement on ranking of search results has also been shown to be quite low due to subjectivity in human judgements (Bar-Ilan and Levene, 2011; Teevan et al., 2007). Despite much research in the information retrieval literature on ranking, for example, the well-known TF-IDF (Salton, 1989) based on processing the text in documents, or the PageRank (Brin and Page, 1998) based on the hyper-textual structure of the web, it is not so clear how search engines incorporate popularity and individual preferences into the ranking formula. In particular, it is not well-understood to what extent user preferences change in time and how such knowledge may inform search engines' rankings.

In this study we investigate an under-explored topic in the literature: change of user assessment of search results over time. Previous work thoroughly 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. These changes are explained by the fact that users understand their needs better in the end rather than in the beginning of search process, and try to refine their search to get the optimal results. In this study we explore a different type/dimension of change in users' evaluation of relevance, i.e. the "change in time". This change if it exists reflects the essential inherent subjectivity of users' perception and evaluation 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 we ask users 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, say, a few weeks or months? It is worth mentioning that in any case some latent factors that affect users judgements might exist, such as learning (Serola & Vakkari, 2005; Vakkari, 2016) or change of interest, but these are difficult to measure.

Therefore, the major goal of this research is to investigate *how*, i.e. in what manner, do users' preferences change and to understand the nature and identify patterns of these changes (rather than *why* the changes occur). As opposed to previous work (Scholer, Turpin and Sanderson, 2011), we believe that such changes are not caused by errors and propose a theoretical model to explain these changes. We borrow from the general theory of modelling user preferences into a finite set of "coarse categories"

(Mullainathan, 2000), where individuals tend to categorise objects into a coarse rather than a fine set of categories. Coarse categorisation leads individuals to “coarse beliefs” (Bjorndahl et al., 2013), where inferences are based on a finite set of probabilities rather than a continuum. A coarse thinker is unable to distinguish between situations within the same category and thus applies the same model of inference to all the situations in the category. Examples of coarse categorisation are the star ratings given to hotels and products, and, in the context of this paper, the relevance judgements attached to search results. In particular, we assume that users tend to distinguish only between a small number of relevance categories and therefore group results according to these coarse categories. Moreover, users find it easier to attach a relevance judgement to a search result, as opposed to assigning it a precise rank, which is a much finer value judgement. In addition, coarse beliefs often seem more natural than fine-grained beliefs when it comes to modelling human preferences, as shown by several examples in (Bjorndahl et al., 2013). Regarding search engine results, users cannot generally distinguish between results that fall into the same coarse category and thus may swap their assigned ranks between a first round of ranking and a second one occurring at a later time. Thus, this type of “local” inside-category changes does not reflect any change in user opinion regarding the judged search result. Hence, another goal of this research is to estimate what part of the users’ ranking changes are local, and, on the other hand, what part of the changes reflect real changes in user opinions about the result preference.

To address the above questions we conducted a large-scale study with 86 human subjects. The subjects had to rank and assess the relevance of an identical set of search results (presented in a different order) twice within a two months interval. To quantitatively evaluate the changes in users’ preferences of the produced rankings and reveal the locality and coarseness patterns in these changes, we propose two new measures: the *change coefficient* and the *change in k-subset* of ranks. These measures consider the proportion of change at different locality levels and also enable to count the intra vs. inter-relevance category changes of different types. To assess whether users apply categorical thinking when doing relevance judgments and rankings, we would expect less inter-category changes than intra-category changes, as calculated by the above measures, using repeated measures ANOVA with SPSS (Field, 2013).

We found that the majority (61-70%) of changes are swaps of ranks within less than 3 places (for ranking the top-ten results out of 20), and for relevance judgments 85-90% were within distance of 1 (when 4 possible relevance judgements are possible). Hence, our results provide evidence for existence of coarse categories and categorical thinking of users when judging the relevance of search results.

Relevance is a central notion in information science and is an important part of user information seeking models (Saracevic, 2007). As further stated by (Saracevic, 2007, p. 2139): “The role of research is to make relevance complexity more comprehensible formally and possibly even more predictable”. Human evaluation of documents relevance is needed in many fields and purposes so it is important to understand the factors and phenomena behind it. In this broad context/framework, our research contributes to understanding and modelling the change in time in user relevance evaluation behaviour.

The problem is also important in understanding whether and to what extent there is a need for personalization and adaptation of the search engine's ranking over time.

The rest of this paper is organized as follows. In the next section we review the related work, then we describe our study setup, further our results are presented and discussed and finally, we provide some conclusions and future research directions.

Related work

Change in user assessment of search results is an under-explored area of research. Here we review the most relevant user studies related to agreement on ranking and relevance judgements. 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 all these experiments the result set was based on the top-10 or top-20 results of the search engines for a given query. These studies only asked the users to rank the results, without asking for their relevance judgements, and the users were asked to rank the results only once. In (Hariri, 2011) the authors also studied Google rankings, and asked users whether they considered the top results to be more relevant. The study was based on the search results of 34 different queries and the results were judged by the user submitting the query. Suprisingly, the fifth ranked result was judged to have the highest relevance, slightly more than the top ranked result.

Saracevic (2007) reviews the existing research on relevance evaluation and its dynamics. The reviewed research examines the change in the criteria of relevance assessment and factors that influence these criteria as the search process evolves. The first work on studying dynamic changes was by Rees and Schultz (1967). Other studies based on more than two subjects include (Vakkari and Hakala, 2000, Vakkari, 2001 and Tang and Solomon, 2001). These studies report that relevance criteria remain quite similar among users and for the same user in different search iterations, however their weights might change. Serola & Vakkari (2005) conducted a user study with 22 psychological students comprising two sessions of evaluation of search bibliography for proposals' expectations and assessed contributions. They report that at the beginning of their task, the students knew so little about the subject that they based their assessment more on the aboutness, i.e., on the general topicality of the references. At the second session, after they had learned more about their topic, the students became more open to accepting other types of information than they expected and could more easily change their goals during the search session and assess how to use the information types provided.

In contrast, in our study the participants were explicitly instructed to use the same evaluation criteria and not change their criteria and goals in the second round, with respect to the same queries and result sets. Also, here we explicitly concentrate on the case of web search, and examine what happens for two separate standalone search sessions when the task and the goals are identical, and assessed at two different points in time. The only difference in the setting between the two sessions is that in the second session 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, Turpin and Sanderson studied repeated relevance judgements of TREC evaluators (Scholer et al., 2011). 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, we measure changes in users' rather than experts' judgements, and also for ranking of results rather than just for binary relevance judgments. Our experimental setting is different too, as in our case the judged documents are search engine results from the web.

Scholer, Kelly and Webber (2013) asked their users to evaluate the relevance (on a 4-point scale) of 3 documents twice, as part of a larger scale experiment. The “temporal” aspect was judged by inserting “duplicate” results within a single session. The reported self-agreement was only about 50%. In this study the time interval between the two evaluations was less than one hour as they were part of the same experiment; the authors do not interpret this result in their paper.

In (Ruthven et al., 2007) the authors conducted a two-round experiment of answer assessments to the same TREC Question Answering session, considering top-5 answers to 30 selected questions. Each answer consisted of a text fragment and was assessed according to the presence of nuggets - facts or concepts relevant to answering the question. To assess the consistency of the assessments a measure was devised to calculate the overlap between the two sets of assessment. The overlap values ranged from 0.95 to 0.61 with a mean assessor overlap of 0.85. The task of question answering is different from search, since both the questions and the answers are much more focused and narrow-formulated, and are thus assumed to be easier to judge. The setting of the study differed from ours as well: the scale was binary, only top-5 results were assessed for each query; there was no analysis of the change patterns, and the changes in judgments were interpreted by the authors as inconsistencies and errors.

Another somewhat related experiment was arranged by Sormunen (2002). Two rounds of reassessment of TREC-7 and TREC-8 documents with a 4-point relevance scale were performed. However, in a second round of assessment the users were intentionally exposed to their first round judgments and the TREC’s original judgments for the documents, in order to influence their final decision. The analysis of the changes demonstrated that in the majority of the cases the users did not change their judgments. Most of the changes were for the irrelevant documents. The ambiguity of the query topic was detected as the main reason for inconsistencies. As opposed to our setting, in this study there was a direct influence of the previous judgments on the changes in the second round's decisions. In addition, in our case both search topics were clearly defined and non-ambigious.

In summary, the main differences between our study setting and the settings of the previous studies above are that: 1) in our experiment the participants performed two identical tasks of independent assessement rounds, 2) and produced two types of evaluation: relevance judgments of 20 results for every query on a scale of 4 (rather than 2 or 3-point scale as in the most of previous work) and rankings on a scale of 10. The latter one (rankings), to the best of our knowledge has not been examined for change in any previous research. It is well-known that relevance is subjective and situation dependent (Saracevic, 1996; Mizzaro, 1998), and therefore a document judged to be relevant in a certain situation and time might not be judged to be relevant later on. Therefore, we believe that the changes in ranking and relevance judgements are not necessarily errors (as argued in (Scholer et al., 2011)), and aim to analyse the patterns of these changes.

Methods

In this section we describe the methodology applied to examine the change patterns in users' evaluation of search results' relevance and rankings over time.

Study Setup

Eighty six information science students were randomly assigned to one of two groups. All the students were enrolled to the "Introduction to information science" class. In general it is preferred that users search on topics of their interest, but in this case the results cannot be aggregated, and thus the compromise is to define a scenario, provide the users in the group with an identical set of results, and ask them to carry out exactly the same task.

There were two topics, Big Data and Alzheimer. For each topic the participants were given a randomly ordered list of 20 search results for the given query ("Big Data" and "Alzheimer"). The search results were presented in a similar style to SERPs, i.e. title, URL and snippet as displayed by the search engine.

Students were presented with the following scenario: "Your aim is to learn about the given topic based only on the search results, in order to be able to prepare a good summary of the topic" (they did not actually have to submit the summary). The students were guided that the order of the items in the list presented to them has no relation to their relevance to the query. We were motivated by the study of Spink & Greisdorf (2001), who assert that there is no consensus regarding how relevance judgments should be measured and thus no specific instructions on this matter can be provided to the users. We also adopted Spink & Greisdorf's (2001) 4-point scale for relevance judgments. Hence we asked the students to assess the relevance of all 20 items as one of the following: not relevant (1), slightly relevant (2), somewhat relevant (3) and relevant (4).

In addition, the students were asked to rank the 10 best results in their opinion, where no ties were allowed. There were two reasons we decided to ask users to rank only the top 10: 1) search engines present 10 results as the default setting, and 2) it requires considerable effort to sift through 20 results and to rank them without ties. Google Forms were used to collect the answers. The participants were asked to repeat the same exercise two months later. The query topics were not part of the curriculum and were not studied either in this class or in any other classes the participants took, but of course they might have been (and probably were) influenced by what they learned when they assessed the results for the first time.

The search results displayed to the participants were collected from Google and Bing, and there were two versions of search results for each query. The first set comprised the top results from Google and Bing (top-10 Google and top-10 Bing, supplemented, because of the partial overlap between the results). The second set comprised the Google results displayed on the first and the tenth result page (i.e. results 1 to 10 and 101 to 110). The Big Data query was submitted in English and the Alzheimer query in Hebrew, the mother tongue of most of the participants. The 86 participants were randomly assigned to one of two groups of approximately the same size (over 40 in each group). The first group was shown the Google-Bing (called *Google&Bing*) results for Alzheimer, and the first and tenth page results from Google (called *Google10&100*) for Big Data, while the second group was shown the first and tenth page results from Google for Alzheimer, and the Google-Bing results for Big Data. Thus, there were four different tasks with two queries and two result sets for each query.

The second time the participants were given exactly the same instructions and evaluation criteria, and exactly the same set of queries and search results. We stressed that there is no "correct answer" in relevance and ranking judgments, and the students were explicitly asked to not to consult or view their previous judgments even if they have saved them. In addition, to prevent the participants from using or recalling their previous assessments, the results in the second round were displayed in a different random order than in the first round. In both rounds all the participants in a given group saw the results in the same random order. 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.

It is known that presentation bias can influence users' preferences (Joachims et al., 2007). However, it has been shown that when the number of results is small then the order of their presentation does not affect the ranking (Saracevic, 2007). We decided that reordering the second time would be more beneficial, since they make it harder for the students to re-use their previous judgments. Although, the rankings might have been influenced by the different display order the second time, this was not the case in our experiment as for three out of the four tasks there was no correspondence between the order of the results' representation to the users and their ranks as shown in Table 1. Only for the Alzheimer *Google&Bing* task two of the first displayed results were ranked at the top-3 ranks by the users (see Table 1). Further analysis shows that this task has some features that distinguish it from the other three tasks. These features are discussed in more detail in the results and discussion section of the paper.

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**	**Ranke d in Round 2 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 ranking and relevance judgment tasks were part of the participants' class assignment. They 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 class instructor by email. They were also told that this decision had no effect whatsoever on their grades. No students asked to withdraw their data. Although the experiment involves human subjects (students), no personal information was gathered on them. The Faculty of Humanities' IRB (ethics committee) waived the need for written consent. There were no minors enrolled in the study. The IRB of the Faculty of Humanities at Bar-Ilan University approved the experiment.

Patterns of changes in user evaluation of search results

Our goal is to investigate how users' evaluation changes in time.

In the introduction we have already mentioned the theory of "coarse categories" (Mullainathan, 2000) and "coarse beliefs" (Bjorndahl et al., 2013). In particular, we hypothesise that each of the four relevance categories may be considered as a coarse category, and that subsets of ranks such as the top-5 and last-5 ranked results may also be considered as coarse categories. If this hypothesis is true, then changes in user assessment of relevance and ranking will be primarily local, showing the difficulty that users have, for example, of deciding whether a search result should be ranked at position 4 or 5. Moreover, when a change of category occurs, it is primarily between neighbouring categories, for example, a user may decide to change their assessment of a search result from "relevant" to "somewhat relevant". Mullainathan (2000) formalised categorical thinking using a Bayesian model, where people hold only a finite number of posterior beliefs. However, in our scenario of users assessing search results over time, we simply wish to quantify the changes in user behaviour over time according to the following two change patterns:

1) *Coarseness* – according to this pattern users distinguish between a few coarse categories of relevance (following the principle of "categorical thinking" advocated by (Mullainathan, 2000)) and do not perceive relevance as a continuous fine-grained range of values. This implies that, results are grouped according to these categories, where all the results inside a category are evaluated as having the same relevance.
2) *Locality* – this pattern holds if users do not distinguish between closely ranked results or results in the same relevance category, and tend to swap their ranks (in accordance with (Mullainathan, 2000) stating that a categorical thinker cannot sufficiently distinguish between the “types” in a category).

As our relevance evaluation allows for ties, it actually creates groups of results according to their relevance grade. Therefore, the first kind of coarse categories we define and use in this study is based on the relevance grades, which seems to be natural. Ranking, in contrast, represents a more fine-grained scale of values. Hence, a second kind of coarse categories that is examined in this research is of subsets of ranks of size *k*, for example the top-5 and last-5 subsets. We further test the existence of the above patterns in our data, which may indicate that some of the changes are indeed local, as mentioned in the introduction.

Measures of change in user evaluation of search results

To answer the above questions we first investigate how the results are distributed with respect to the relevance categories, i.e. whether some categories contain more results than the others, and whether results with higher user-assigned ranks (recalling that the highest rank is 1 and the highest relevance value is 4) are concentrated mostly in higher user-assigned relevance categories. To this end, for both rounds of the experiment we count the number of results in each relevance category for all the results, for the top-10 results only, and for the last-10 results only.

Patterns of locality and coarseness can be quantified by measuring changes at varying distances between the relevance judgements or assigned rankings in round one and round two. In particular, the differences between changes for each result with respect to ranking versus the corresponding changes in relevance judgements can show the influence of the evaluation scale on the *change coefficient* defined below.

Hence, for each query and result set, we assess what proportion of the results in the set was *not* given identical ranks or relevance judgments by a specified user, on the first and second rounds of the experiment. This measures the amount of change at the *exact match* level (i.e. we consider results with distance 0 between the two ranks or relevance values as identically judged). Further, we consider the case when the rankings or relevance judgements were not precisely identical in both rounds but still sufficiently close, or in other words fall within the same coarse category.

We now define a new measure for the amount of change in the results of a given coarse category. For each category, we consider the results that were assigned to a category in at least one of the evaluation rounds of the experiment. Formally, we define the *non-local change coefficient* for a category, c, at a distance, d, with $0 \leq d \leq |c|$, for a given set of results, $s1, s2, \ldots, s_i$, evaluated twice by a user, u, either with ranking or relevance values, $r1(s)$ and $r2(s)$ for a result s, as follows:

$$\Omega(c,d) = 1 - \frac{\sum_{i=1}^{|c|} n(s_i)}{|c|}, \text{ where } n(s_i) = \begin{cases} 1, |r1(s_i) - r2(s_i)| \leq d \\ 0, otherwise \end{cases}$$

Thus, $\Omega(c,d)$ is the proportion of the results that are within distance d in category c only once, i.e. either in the first or the second round they were assigned to a category, at distance larger than d from c. In our experiments this measure was applied on the first kind of coarse categories, naturally formed by the relevance grades (1-4) assigned to the results in the two rounds of the experiment.

Then, based on the above definition we define the *global change coefficient* for the entire set of query results, R, when c is set to R. In this case we denote it as $\Omega(d)$, which is a shorthand for $\Omega(c,d)$, when all the results are in a single category, i.e. it is the proportion of results in R that are at distance greater than d in the two rounds. 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 R 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 higher than 11 are considered. This is because results that were assigned rank 11 were not actually ranked by the users.

In addition, for the second kind of coarse categories, based on subsets of k ranks, we measure the proportion of new *non-overlapping* results in the subset of k ranks starting at a position p, that were introduced in the second round of the experiment. More formally, given a set of results R, and a consecutive subset of ranks $(r_p, r_{p+1}, \ldots, r_{p+k})$ where $1<r_i\leq|R|$, we construct two subsets of ranked results with the corresponding ranks for each of the two rounds, $R1$ and $R2$. We define the *change in k-subset* measure, as follows*:*

$\Psi(p, k) = 1 - |R1 \cap R2|/k$, for some p and $k=\{1..N\}$

In the sequel $\Psi(top\text{-}k)$ will stand for $\Psi(1,k)$ and $\Psi(last\text{-}k)$ will stand for $\Psi(N\text{-}k+1,k)$. This measure computes the proportion of non-local changes a user has made in a defined subset of ranks of size k. It complements $\Omega(c,d)$, since $\Psi(p, k)$ is applied to the ranked-based coarse categories, while $\Omega(d)$ is applied to the relevance-based coarse categories. To illustrate the calculation of the above measures, in Tables 2 and 3 we present an example from our data set for one query and one user.

Table 2. An example of ranking and relevance values in the two rounds for one user and one query. Recall that category 4 for relevance stands for the most relevant results to the query.

Result number	**Ranking round 1**	**Ranking round 2**	**Relevance round 1**	**Relevance round 2**
1	11	10	1	2
2	6	4	3	4
3	4	2	4	4
4	10	6	2	3
5	11	11	1	1
6	2	5	4	3
7	7	11	2	1
8	3	3	4	4
9	11	11	1	1
10	8	11	2	1
11	5	7	3	3
12	11	11	1	1
13	1	1	4	4
14	11	11	1	1
15	9	11	2	1
16	11	11	1	1
17	11	9	1	2
18	11	8	1	2
19	11	11	1	1
20	11	11	1	1

Table 3. The results of the calculation of Ω for the above data for the four categories based on rankings and relevance judgements for distances of 0 and 1, respectively.

Ranking - change coefficient per category	**Relevance - Change coefficient per category**
Ω(1,1)=0.31	Ω(1,0)=0.46
Ω(2,1)=0.86	Ω(2,0)=1.00
Ω(3,1)=1.00	Ω(3,0)=0.75
Ω(4,1)=0.60	Ω(4,0)=0.40

For example, for the ranking of category 1 and distance 1 we consider the 13 out of the 20 results that were judged with relevance 1 on at least one of the rounds (numbers 1, 5, 7, 9, 10, 12, 14-20). Only results 7, 15, 17 and 18 have different ranks in each of the rounds with distance greater than 1. Thus, $\Omega(1,1)=4/13=0.31$.

The global change coefficient for relevance with distance 0, $\Omega(0)$, is 0.45 (9 out of 20), with distance 1, $\Omega(1)=0$, and for ranking it is $\Omega(0)=0.85$ (11 out of 13 results – 7 were ranked 11 both times and are thus removed from this calculation, and the rankings of results 8 and 13 were identical on both rounds) and $\Omega(1)=0.77$ (10 out of 13), respectively. Hence, as distance increases the change coefficient decreases. In addition, for ranking the change in *k*-subset for the top-10 and top-5 are, *Ψ(top-10)* is 0.3 (3 out of 10 as results 2, 3, 4, 6, 8, 11, 13 were ranked in top-10 in both rounds) and *Ψ(top-5)* is 0.2 (4 out of 5 top results: 3, 6, 8, 13 were overlapping).

The change coefficient defined above can be viewed as a finer grained measure than the known rank-correlation measures used to compare ranked lists (Fagin, Kumar & Sivakumarm, 2003). So, for example, the correlation between the two lists would be one if the change was zero for d=0. In particular, the change coefficient allows us to detect at what distance a change occurs and the proportion of results that are involved in the change. Moreover, the distance is a locality measure that will help us in testing our hypotheses regarding coarse categories. In addition, the *non-local change coefficient* for a given category is particularly suited to measure and compare intra-category changes making this evaluation transparent.

In addition to the evaluation of the change patterns, we wish to investigate whether some users tend to make fewer changes than others independently of the query and result set. To this end, we perform direct user behaviour analysis by computing the Pearson’s correlation between the proportion of changes (the change coefficient values) in judgements of the same user in different tasks. (Recall that every user has participated in two tasks and has judged two different sets of results for different queries.)

In summary, we propose a new methodology to assess the influence of the temporal factor on change in user relevance judgments and rankings. To this end, we have carried out a large-scale two-round experiment, where the same groups of subjects had to accomplish the same evaluation tasks twice within an interval of a few weeks. The existence of two types of change patterns, adopted from the theory of categorical thinking, is investigated in this paper: locality and coarseness. To measure the level of locality the global change coefficient measure was employed, and to measure the level of coarseness for the 4-scale relevance judgments the non-local change coefficient for a category was devised, while to measure the level of coarseness for the 10-scale rankings the change in k-subset ranks metric was used.

Results and Discussion

Pattern analysis for individual user judgements

As a first step to explore the coarseness patterns in the data, we calculated the distribution of the results within the four relevance grades for each user, and then averaged the results over all the users. Each column in Figures 1-3 show the average number of results within a corresponding relevance grade for a given user group and round.

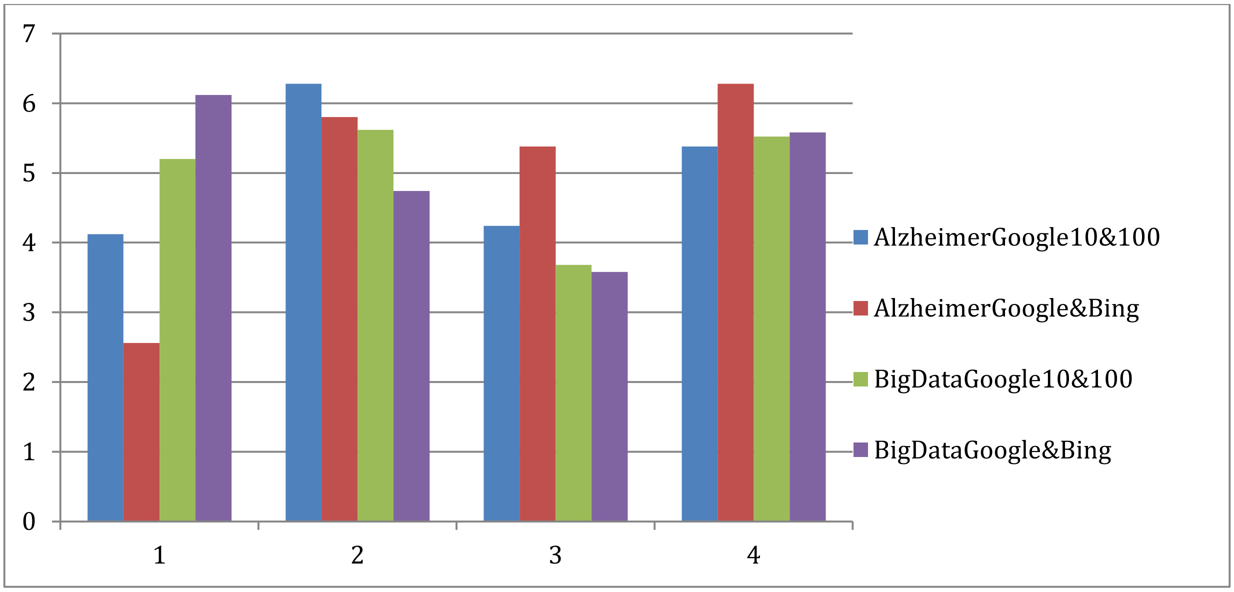

Figure 1: The distribution of the 20 results within the four relevance grades by the user rank levels for the first round of the experiment. Recall that category 4 was assigned to the most relevant results for the query, while category 1 was for the least relevant results.

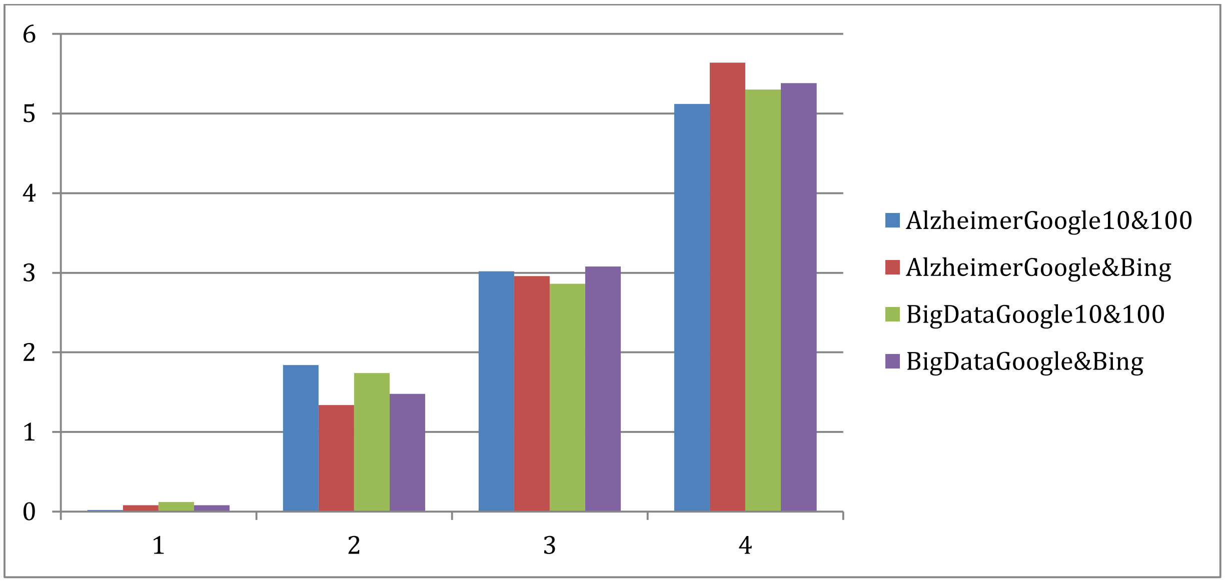

Figure 2: The distribution of the top-10 results within the four relevance grades by the user rank levels for the first round of the experiment.

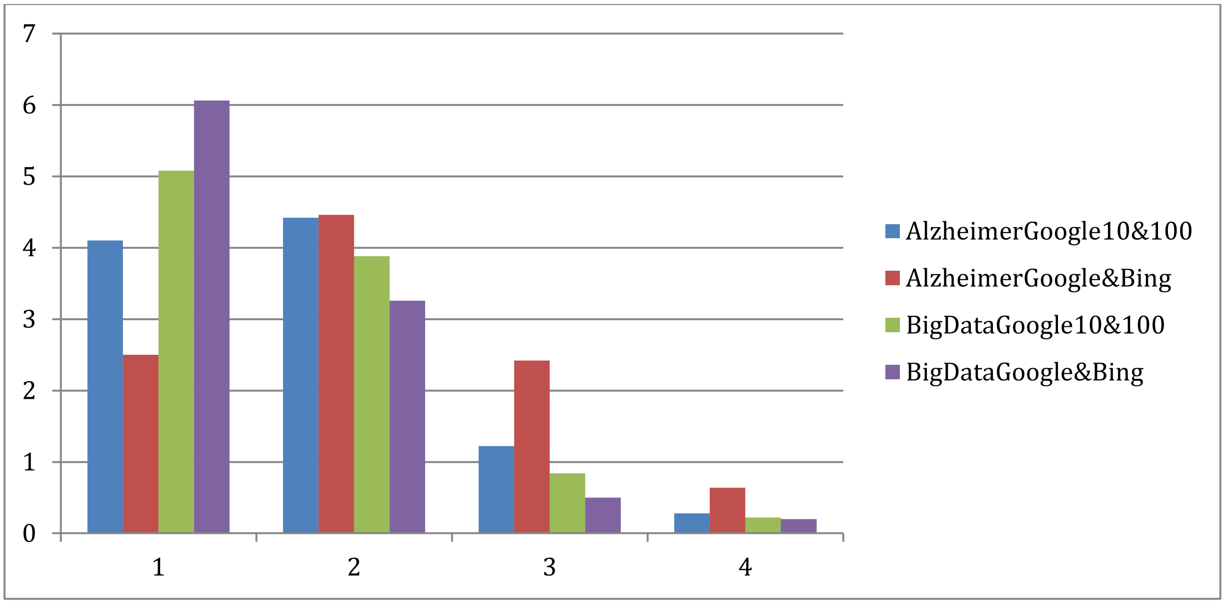

Figure 3: The distribution of the last-10 results in the four relevance categories by the user rank levels for the first round of the experiment.

As can be seen from Figure 1, overall, when taking into account the top-20 ranked results, the relevance judgements are, more or less, evenly distributed between the four grades of relevance. We can then see in Figure 2 that the top-10 ranked results are concentrated within the top-2 relevance grades (4, 3), while in Figure 3 the non-ranked last-10 results are mostly concentrated in the last two relevance grades (1, 2). We also notice from Figure 1 that for the Alzheimer Google&Bing task, the number of results in relevance grade 1 (not-relevant) was much lower than for the other tasks within this grade, while for relevance grades 3 and 4 the amounts of results for this task were much higher than for the others. This might explain the presentation bias found for this task as shown in the previous section, since for this task the top presented results in both rounds were indeed quite relevant to the query. For the second round of the experiment the distributions were quite similar to those of the first round.

Next, we investigated the changes of individual users' rankings and relevance judgements for the same 20 results between the first and second rounds of the experiment. To this end, we calculated the average of the global change coefficients, $\Omega(d)$, with *d*=0 over the individual users' rankings and relevance judgements. The results for different tasks are presented in Table 4. We observe that, in general, similar values were obtained for the different queries and result sets, recalling that when $\Omega(d)$=0 then no change occurred. We do not show the change coefficient for relevance grades with *d>1*, since they are very close or equal 0; these are denoted by N/A in Table 4.

Table 4. The average global change coefficient values for ranking and relevance with different distances and result sets. Standard deviation values are shown in parentheses following the average. *Rel.* stands for relevance.

	AlzheimerGoogle10&100 Ranking N=45	**AlzheimerGoogle10&100 Rel.** N=45	**Alzheimer Google&Bing Ranking** N=41	**Alzheimer Google&Bing Rel.** N=41	**BigData Google& Bing Ranking** N=42	**BigData Google& Bing Rel.** N=42	**BigData Google10 &100 Ranking** N=41	**BigData Google10 &100 Rel.** N=41
Ω(0)	0.87(0.18)	0.48(0.18)	0.87(0.18)	0.53(0.17)	0.84(0.17)	0.52(0.17)	0.87(0.17)	0.43(0.14)
Ω(1)	0.61(0.23)	0.12(0.10)	0.67(0.21)	0.15(0.13)	0.62(0.24)	0.13(0.12)	0.64(0.24)	0.10(0.10)
Ω(2)	0.45(0.22)	N/A	0.50(0.20)	N/A	0.44(0.23)	N/A	0.48(0.23)	N/A
Ω(3)	0.33(0.19)	N/A	0.39(0.17)	N/A	0.31(0.20)	N/A	0.32(0.19)	N/A

Then, to assess the frequency of the locality patterns, the global change coefficient was measured for distances greater than or equal to one between judgements in the two rounds. We experimented with $d=1$ for relevance, and $1 \leq d \leq 3$ for rankings (as the distance between judgements of the results in the two rounds), also presented in Table 4.

In order to test the significance of the decrease of the change coefficients we report the F-statistic tests from a repeated measures ANOVA (used to detect any overall differences between related, non-independent means). All the tests were significant as can be seen in Table 4a. It should be noted that in the first group there were 45 participants, but only 42 of them ranked and assessed the relevance of their second query: BigData Google&Bing. We also tested pairwise significance for ranking, and all the pairs were significantly different with high eta^2 (the effect size).

Table 4a. F-tests results of the repeated measures ANOVA.

Alzheimer Google10&100 N=45	Ranking	F(3,132)=233.98, p<.001, eta^2=0.84
	Relevance	F(1,44)=287.88, p<.001, eta^2=.87
Alzheimer Google&Bing N=41	Ranking	F(3,120)=233.42 p<.001, eta^2=0.85
	Relevance	F(1,40)=258.47, p<.001, eta^2=.87
BigData Google&Bing N=42	Ranking	F(3,123)=245.62, p<.001, eta^2=0.86
	Relevance	F(1,41)=264.00, p<.001, eta^2=.87
BigData Google10&100 N=41	Ranking	F(3,120)=219.58 p<.001, eta^2=0.85
	Relevance	F(1,40)=385.97, p<.001, eta^2=.91

As expected, we can see a significant decrease in the change coefficient, especially the decrease between distance 0 and distance 1, for the relevance judgements, which indicates that most of the changes in relevance judgements were local between the rounds. It also provides some evidence that users are evaluating the results within four coarse categories pertaining to the relevance grades. Users do however change their minds about their relevance judgements, these changes being mostly local, since we are considering the change coefficient at distance 1; the changes are on average 12.5% in this case (see **Ω**(1) for relevance in Table 4). We assume that the number of non-local changes within each coarse category should be as low as possible, which is supported by the low value of the change coefficient for relevancies at distance 1. We further computed the change coefficient individually for each of the said coarse categories pertaining to the four relevance grades. The division of the results into four coarse categories according to their relevance grades seems to be the most natural, as argued above when inspecting the results shown in Table 4. In particular, we computed the change coefficient for each category separately by considering distances between the relevance grades, as shown in Table 5.

Table 5. The proportions of non-local changes for relevance judgements for four coarse categories and various distances. Standard deviation values are shown in parentheses following the average.

	Ω (1,0)	**Ω (2,0)**	**Ω (3,0)**	**Ω (4,0)**
AlzheimerGoogle10&100	0.38(0.34)	0.57(0.29)	0.64(0.28)	0.36(0.31)
AlzheimerGoogle&Bing	0.43(0.42)	0.62(0.28)	0.64(0.32)	0.43(0.26)
BigDataGoogle10&100	0.45(0.32)	0.61(0.26)	0.62(0.32)	0.38(0.32)
BigDataGoogle&Bing	0.26(0.24)	0.53(0.28)	0.63(0.25)	0.41(0.25)
Average	0.38(0.33)	0.58(0.28)	0.64(0.29)	0.40(0.29)
	Ω (1,1)	**Ω (2,1)**	**Ω (3,1)**	**Ω (4,1)**
AlzheimerGoogle10&100	0.16(0.18)	0.13(0.14)	0.13(0.20)	0.24(0.20)
AlzheimerGoogle&Bing	0.34(0.35)	0.21(0.26)	0.08(0.13)	0.23(0.20)
BigDataGoogle10&100	0.20(0.23)	0.13(0.14)	0.16(0.18)	0.23(0.22)
BigDataGoogle&Bing	0.12(0.13)	0.12(0.13)	0.12(0.17)	0.20(0.18)
Average	0.21(0.22)	0.15(0.17)	0.12(0.17)	0.22(0.20)

The results in Table 5 for changes in relevancies are similar, on average, to the global changes for relevancies, shown in Table 4. However, Table 5 gives us a finer picture than Table 4, and we can see that for *d=0,* for the most relevant and not relevant results (categories 4 and 1), the non-local changes were minority and were less than for the somewhat relevant results (categories 2 and 3). Also, in all cases, the results in Table 5 for the extreme categories are considerably lower than for the middle categories. This indicates that categorical thinking phenomenon holds mainly for the extreme categories. Here too we applied repeated measures ANOVA, for *d=0,* non local changes for the four relevance categories, and the differences between categories 1 and 4 versus 2 and 3 were significant at level *p<.05* for all pairs (1-2,1-3,4-2,4-3) for all four queries.

The reasons for proposing four coarse categories are that each category naturally corresponds to a relevance grade, and also having an even distribution of results among the four relevance grades. Although we have already presented evidence for four coarse categories, we further checked an alternative hypothesis that there are actually only three coarse categories obtained by merging adjacent relevance grades; the three possible divisions into three categories as thus:{{1,2}, {3}, {4}}, {{1},{2,3},{4}} and {{1},{2},{3,4}}. We then calculated the average change coefficient for relevancies of these possibilities within each of these categories, as shown in Table 6. It can be seen that these results are slightly lower yet compatible with those for four categories, as shown in Table 5, and thus there is no reason to prefer a coarser division into three relevance grades by merging adjacent grades from the four original ones.

Table 6. The average change coefficient values for all three possible divisions above of the three coarse categories with distance 0 for relevance judgements. Standard deviation values are shown in parentheses following the average.

	Ω (1,0)	**Ω (2,0)**	**Ω (3,0)**
AlzheimerGoogle10&100	0.37(0.28)	0.51(0.26)	0.33(0.27)
AlzheimerGoogle&Bing	0.54(0.36)	0.56(0.28)	0.36(0.23)
BigDataGoogle10&100	0.25(0.19)	0.52(0.27)	0.33(0.27)
BigDataGoogle&Bing	0.22(0.20)	0.51(0.24)	0.36(0.21)
Average	0.35(0.26)	0.53(0.26)	0.35(0.25)

Next, we computed the change for the second kind of coarse categories defined for ranking, based on the subsets of *k* ranks. Table 7 considers the change in the top-5, last-5 and unranked result subsets. We chose *k=5* creating two subsets for the top-10 ranked results (*top-5* and *last-5*), since these subsets roughly corresponds to the two highest relevance grades, which cover most of the results ranked by the subjects, as shown in Fig 2. We then measured the non-local changes in the top-5, i.e. *Ψ(top-5),* and last-5, i.e. *Ψ(last-5),* subsets and in all the unranked results, i.e. *Ψ(unranked),* the results that were unranked at least once, as a third coarse category for ranks.

Table 7. The change in ranking for *k=11* (all the unranked results), in top-*k* and last-*k* for *k=5* for the different tasks. Standard deviation values are shown in parentheses following the average.

	Ψ(unranked)	***Ψ(last-5)***	***Ψ(top-5)***
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(0.23)

Interestingly, here too there is significantly 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 one, for which the majority of the results are local within the category. We applied repeated measures ANOVA, for non local changes for the three rannking categories, and the differences between categories top-5 and unranked versus last-5 were significant with $p<.001$ for all pairs for all four queries.

The obtained results indicate that for both types of coarse categories (ranking and relevance-based) there is a strong evidence for categorical thinking especially for the extreme categories. The largest amount of non-local changes in most cases was found for the AlzheimerGoogle&Bing task. The reason for this behaviour might be the highest number of relevant results as discussed earlier.

User internal consistency analysis

Finally, our aim was to examine, across both queries, whether some users are more consistent in their judgements than the others. Recall that every user judged two diverse result sets (of two different queries). Thus, for the consistency test, the exact match change coefficients $\Omega(0)$ for the two result sets assessed by the same user were computed and then compared by the Pearson's correlation. We found quite a strong correlation of 0.58 (t-test significant at $p<0.01$) on average over all users, which means that there is a strong linear relationship between the change coefficient values independent of task and query. In other words, some users are always more stable in their evaluation, while others tend to make more changes in their judgements of search results, however the level of the internal user consistency among different result sets is quite high.

Discussion and conclusions

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.

The primary goal of this study was to investigate how the user preferences change in time, by assessing the internal change patterns of users' ranking and relevance judgements of query results. In particular, we distinguished between two change patterns borrowed from the the theory of categorical thinking: (1) coarseness and (2) locality. Moreover, we proposed two new measures of change for ranking and relevance judgements, the non-local change coefficient, $\Omega(c, d)$, for a category *c* at distance *d*, and the change in *k*-subset measure *Ψ(p, k)*, for a consecutive subset *k* of ranks starting at position *p*.

To aid our investigation, we conducted a large-scale user study for ranking and judging the relevance of query result sets, and repeated the experiment within a two-month period. We found that large changes (of about 50% when *d*=0) occur in users' relevance judgments between the two rounds. The figures we obtained are comparable to those of (Scholer et al., 2013), but higher than those of (Ruthven et al., 2007; Scholer et al., 2011; Sormunen, 2002), who reported 15-40% change of relevance assessements in their experiments. The diversity of these findings might be attributed to the differences in the experimental settings of these studies. For rankings, we obtained that 61-70% of the changes were local within *d*=3, while the previous studies did not examine the change in fine-grained rankings over time.

Nevertheless, the majority of the changes for relevance judgements (85-90%) were local (within *d*=1) for different queries and result sets, which provides empirical evidence for using coarse relevance categories. These results on the locality of changes, support and extend the findings of Sormunen (2002). In his experiment most of the changes were for irrelevant documents, where 30% of these documents were reassessed as marginally relevant (a local change with *d*=1) and only 6% of them were reassessed as relevant (non-local change with *d*=2) in the second round.

The change within the relevance categories (especially for the extreme categories) was lower than the global change, both for rank-based and relevance-based categories. Thus, it seems that users are more likely to swap results within a category than to move results in and out of a category, thus providing further evidence for the presence of coarse categories. 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. This result matches and provides an additional evidence to the previous studies findings, such as, (Gwizdka, 2014; Vakkari, Luoma & Pöntinen, 2014). They conducted several eye-tracking experiments and measured the dwell time on documents with different relevance level to the search task. Similarly to our finding that most of the non-local changes occur in the middle relevance categories, these studies showed that users need most time and cognitive effort in assessing somewhat relevant or interesting search results compared to very relevant or non relevant ones. Also, (Sormunen, 2002) showed that only less than 1% of the relevant or highly relevant documents were reassessed differently in the second round of his experiment, hence the author concludes that the assessors were very reliable in identifying relevant and highly relevant documents.

We also examined whether three coarse categories for relevance may be better than four, but found little evidence for this. What the "optimal" number of relevance grades should be, is still an open problem for further research, although we have evidence that three or four is a sensible choice, which is also inline with the findings of previous work (Spink & Greisdorf, 2001; Sormunen, 2002; Gwizdka, 2014; Vakkari, Luoma & Pöntinen, 2014).

The above tendencies were quite similar for all the queries and result sets. In addition, we found that some users tend to change their mind more than others independently of the query and result set.

The main limitation of this study is that it does not consider additional factors, such as learning (Vakkari, 2016) and change of interest, that might have influenced the users' judgments and caused part of the changes. To determine the influence of learning on change of relevance assessment, we conducted a preliminary 3-round experiment (Zhitomirsky-Geffet, Bar-Ilan & Levene, 2015), where the same group of 35 subjects were asked to evaluate the relevance and rank the same 20 results of the same query for three times within a few weeks interval between the rounds. Although one would expect that most of the learning occurs between the first and the second rounds, the results showed that a similar amount of change was measured between the first and second rounds and between the second and third rounds. The users' posteriori explanations of their evaluation change across the rounds revealed that among the most predominant factors were categorical thinking and learning (knowledge acquisition), which according to their reports took place in both initial rounds. A further more comprehensive investigation of the additional factors that influence change in relevance and ranking judgments are a subject for future research.

In summary, this research provides some theoretical ground supported by empirical evidence to modelling relevance evaluation and its patterns of change in time. We highlight the fact that, to the best of our knowledge, this is the first experiment that systematically analyses different patterns of change in time in ranking and relevance assessment. From the results, which support the presence of three-four coarse categories, we conclude that, for assessment purposes, employing relevance judgements with a coarse scale is more suitable for human users than fine-grained rankings. Our research might also provide insights for search engines and motivate them to adopt and expand usage of more structured and organized representation of search results (Google infoboxes was a first step in this direction), displaying groups of results in a user-oriented approach, rather than as a long sequential list. Our findings may also have practical implications for improving ranking and personalisation strategies of search engines, based on identifying the results with potential non-local cross-category change in user evaluation as targets for re-assessment and re-ranking.

References

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*, 58(9), 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 Association for Information Science and Technology*, 60(1), 135-149.

Bar-Ilan J & Levene M. (2011). A method to assess search engine results. *Online Information Review*, 35(6), 854-868.

Bjorndahl A, Halpern JY, & Pass R. (2013). Language-based games, theoretical aspects of rationality and knowledge. In: *Proceedings of the 14th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-13)*, January 7-9; Chennai, India. New York: ACM, pp. 39-48.

Brin S & Page L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7), 107-117.

Fagin R., R. Kumar, and D. Sivakumar. (2003). Comparing top-k lists. *SIAM Journal on Discrete Mathematics*, 17, pp. 134-160.

Field, A. (2013). Discovering Statistics using IBM SPSS Statistics Paperback. by SAGE Publications, 4th edition.

Gwizdka, J. (2014). Characterizing relevance with eye-tracking measures. In: *Proceedings of the 5th Information Interaction in Context Symposium*. ACM, New York, NY, 58-67.

Hariri N. (2011). Relevance ranking on Google. Are top ranked results considered more relevant by the users? *Online Information Review*, 35(4), 598-610.

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

Joachims T, Granka L, Pan B, Hembrooke H, Radlinksi F, & Gay G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Transactions on Information Systems*, 25(2), Article 7.

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

Mullainathan S. (2000). Thinking through categories. MIT working paper. Retrieved from www.haas.berkeley.edu/groups/finance/cat3.pdf.

Rees, A.M., & 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.

Ruthven, I. and Azzopardi, L. and Baillie, M. and Bierig, R. and Nicol, E. and Sweeney, S. and Yakici, M. (2007). Intra-Assessor consistency in question answering. In: *Proceedings of The 30th International ACM SIGIR Conference,* 23-27 July 2007, Amsterdam, Netherlands.

Salton G. (1989). *Automatic text processing*. Reading, MA: Addison-Wesley.

Saracevic T. (1996). Relevance reconsidered. In: *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.

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 Association for Information Science and Technology*, 58(13), 2126-2144.

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

Scholer, Kelly & Webber. (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*, New York: ACM, pp. 623-632.

Serola, S. & Vakkari, P. (2005). The anticipated and assessed contribution of information types in references retrieved for preparing a research proposal. *Journal of the American Society for Information Science,* 56(4): 373-381.

Sormunen, E. (2002). Liberal relevance criteria of TREC – counting on negligible documents? In: *Proceedings of the SIGIR 2002*. ACM: New York, 324-330.

Spink, A., & Greisdorf, H. (2001). Regions and levels: Measuring and mapping users' relevance judgments. Journal of the American Society for Information Science and Technology, 52(2), 162-173.

Tang, R. & Solomon, P. (2001). Use of relevance criteria across stages of document evaluation: On the complementarity of experimental and naturalistic studies. *Journal of the Association for Information Science and Technology*, 52, 676–685.

Teevan J, Dumais ST, & Horvitz E. (2007). Characterizing the value of personalizing search. In: *Proceedings of the 30th international ACM SIGIR conference*, July 23-27; Amsterdam, Holland; New York: ACM, 2007, pp. 757-758.

Vakkari, P. & Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. *Journal of Documentation*, 56(5), 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*, 4(3), 295–310.

Vakkari, P., Luoma, A. & Pöntinen, J. (2014b). Book's interest grade and dwell time in metadata in selecting fiction. In: *Proceedings of IIiX'14 conference*. ACM, New York, NY, 28-37.

Vakkari, P. (2016). Searching as learning. *Journal of Information Science*, *42*(1), 7-18.

Zhitomirsky-Geffet M., J. Bar-Ilan, M. Levene. (2015). How and why do users change their assessment of search results over time? *Poster in the Proceedings of the Annual Meeting of the Association for Information Science (ASIS&T),* November, 2015, St. Louise MI, USA.