Categorical relevance judgment

Maayan Zhitomirsky-Geffet
Bar-Ilan University
Ramat-Gan, Israel, 5290002
Maayan.Zhitomirsky-Geffet@biu.ac.il

Judit Bar-Ilan
Bar-Ilan University
Ramat-Gan, Israel, 5290002
Judit.Bar-Ilan@biu.ac.il

Mark Levene
Birkbeck University of London
Malet st, London, UK, NW11 7YB
Mark@dcs.bbk.ac.uk

*Corresponding author

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

Abstract

In this study we aim to explore users' behaviour when assessing search results relevance based on the hypothesis of categorical thinking. In order to investigate how users categorise search engine results, we perform several experiments where users are asked to group a list of 20 search results into a number of categories, while attaching a relevance judgment to each formed category. Moreover, to determine how users change their minds over time, each experiment was repeated three times under the same conditions, with a gap of one month between rounds. The results show that on average users form 4-5 categories. Within each round the size of a category decreases with the relevance of a category. To measure the agreement between the search engine's ranking and the users' relevance judgments, we defined two novel similarity measures, the *average concordance* and the *MinMax swap ratio*. Similarity is shown to be the highest for the third round as the users' opinion stabilises. Qualitative analysis uncovered some interesting points, in particular, that users tended to categorise results by type and reliability of their source, and particularly, found commercial sites less trustworthy, and attached high relevance to Wikipedia when their prior domain knowledge was limited.

Introduction

Previously, in (Zhitomirsky-Geffet, Bar-Ilan and Levene, 2016a; Zhitomirsky-Geffet, Bar-Ilan & Levene, 2016b; Zhitomirsky-Geffet, Bar-Ilan and Levene, 2017), we have demonstrated that when assessing the relevance and rankings of query results, users tend to group results into "coarse categories" (Mullainathan, 2000). This implies that users assess the relevance of results in the same category similarly, without distinguishing between the degree of relevance of results within the same category. It was also found that users tend to preserve their category-based judgment over time. This was tested on a predefined number of relevance values, i.e. four, (1 – "irrelevant", 2 – "partially/slightly relevant", 3 – "quite relevant", 4 – "very relevant"). However, we did not explicitly ask users to classify the results into categories by their degree of relevance to the given query, but only to assign each result a relevance value and a rank. In the current study we continue this line of enquiry by conducting a series of experiments, in several repetitive rounds, where we directly asked two groups of users to classify the results with similar relevance to the query into *relevance categories*, and assign each category a relevance value that assesses the degree of relevance of the assigned category to the query. The number, content and size of the categories are defined by the user.

In an earlier study (Zhitomirsky-Geffet and Daya, 2015) we showed that reranking of search results according to the most prominent and discriminative subtopics for a given query improves the mean average precision of search results. It transpires that users prefer results focused on/associated with certain popular subtopics of the query, and provide lower relevance judgements for the other results. In the current study we explore what characterises these preferred subtopics (or categories as we refer to them here) may have, and how do users (according to what criteria) make their choice.

The main research questions we address here are:

1) Into how many relevance categories do users typically group 20 search engine results they are presented with?
2) What are the sizes of these categories?
3) How do the number, size and content of the categories change over time, and, if so, how does the process of categorisation stabilise?
4) Are there any differences in categorisation and relevance assessment of the results between the experimental rounds and between the queries?
5) How do users group a list of search engine results into categories, how do they define the categories, and how (i.e., by what criteria) do they assess their relevance to the query?

The above questions pertain to the investigation of the categorical thinking hypothesis in the evaluation process of search results relevance from different perspectives. We expect that if this hypothesis is valid, then the number, size and content of relevance categories for a given set of results should not vary too much when repeating the experimental evaluation, and should stabilise after two to three rounds.

The last research question of this study is:

6) Is there a match between users' category-based relevance ranks and Google's ranking of the search results?

This question is of interest to us, as previous work (reviewed in the Related Literature section) demonstrated that there is still a need to reduce the gap between the search engine's ranking and the user's perception of result relevance. More specifically, here we examine whether this gap (detected for an individual result's relevance) still exists at the category level. Furthermore, one of the main practical implications of this research might be improving the ranking and presentation of results by search engines (e.g. Google) by adopting the categorical thinking paradigm. Therefore, in this study we also investigate the relationship between the users' category-based relevance ranks and Google's ranking of these search results.

Related Literature

Theoretical background – categorical thinking

As a theoretical grounding for the proposed model we make use of the theory of "coarse beliefs" (Mullainathan, 2000,) based on the tendency of people to categorise objects into a coarse rather than fine set of categories. Examples of coarse categorisation are the star ratings given to hotels and product prices, and, in the context of this paper, the relevance judgements attached to search results. In particular, we hypothesise that users tend to distinguish between a small number of relevance categories and therefore group results according to these coarse categories. In addition, coarse beliefs often seem more natural than fine-grained beliefs when it comes to modelling human preferences. Regarding search engine results, users cannot generally distinguish between results that fall into the same coarse category, for example users may consider "relevant" results as being part of a coarse category, and thus, will not change their minds concerning the relevance of a results between a first round of relevance assessment and a second one occurring at a later time, despite a small, "local", shift in opinion regarding the relevance of the result. This type of local category changes does not reflect a change in user opinion regarding the judged search result. To this end, the following change patterns were defined and explored (Zhitomirsky-Geffet, Bar-Ilan and Levene, 2017):

1) *Coarseness* – according to this pattern users distinguish between a few coarse categories of relevance (following the principle of "categorical thinking") and do not perceive relevance as a continuous fine-grained range of values. This implies that, results are grouped into these categories, such that all the results inside a category are evaluated as having comparable relevance.
2) *Locality* – this pattern holds when change in user opinion tends to be "local". By this we mean that, for example, that a user is more likely to change his/her relevance judgement from "relevant" to "somewhat relevant" rather than from "relevant" to "not relevant".

To complement the empirical results reported in (Zhitomirsky-Geffet, Bar-Ilan and Levene, 2017), we proposed a Markov chain model as a theoretical basis for measuring the change in users' judgements across three experimental rounds (Zhitomirsky-Geffet, Bar-Ilan and Levene, 2016a). As was shown in that paper, the Markov chain demonstrates that users' opinions stabilise and converge, and that a majority of the changes are local to a neighbouring relevance category. In further work, we tested how aggregate judgements, known as "wisdom of the crowds" change in time (Zhitomirsky-Geffet, Bar-Ilan and Levene, 2016b). Therein, it was found that aggregated judgments are more stable than individual user judgments, yet they are quite different from search engine rankings.

Here we augment the above investigation on the grouping of search results into relevance-based categories, by asking users from the onset to group the results into relevance categories and repeating the process in the course of three experimental rounds. This allows us to address further research questions as formulated above.

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

Here we review the most relevant user studies related to agreement on ranking and relevance judgements. 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 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 that in the previous studies the users' information need changed as the task evolved, while in our experiments users' tasks remain constant over time.

In another 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. Surprisingly, in this study, the fifth ranked result was judged to have the highest relevance, slightly more than the top ranked result.

Serola and Vakkari (2005) conducted a user study with 22 psychology students comprising two sessions (repetitive *rounds* of the same experiment) involving searching a bibliography regarding the expectations of proposals and their 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. During the second evaluation round, after they had learned more about their topic, the students became more open to accepting other types of information than they had originally expected, and could more easily change their goals during the search process and assess how to use the information types provided by the search results.

In contrast, in our study the participants were explicitly instructed to divide the result set into categories. Also, here we explicitly concentrate on the case of web search, and examine what happens, in three separate standalone search rounds, when the task and goals are identical; moreover, the assessments were made at three different points in time. The only differences in the setting between the rounds were different presentation order of the results and that in the second and third rounds the users saw the given set of documents (or their snippets) once and twice before, respectively.

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 assessing the relevance value of one or more categories rather than only allowing 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, Wu, Lee and Webber (2013) asked their users to evaluate the relevance (on a 4-point scale) of three documents twice, as part of a larger scale experiment. The “temporal” aspect of the experiment was judged by inserting “duplicate” results within a single round, and the reported self-agreement was only about 50%. In this particular 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 task, considering the top-5 answers to 30 selected questions. Each answer consisted of a text fragment and was assessed according to the presence of nuggets, i.e. 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 assessments. 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 their answers are much more focused and narrowly formulated, and are thus easier to judge. The setting of the study differed from ours as well: the scale was binary, only the top-5 results were assessed for each query, there was no analysis of the change patterns over time, and the changes in judgments were interpreted by the authors as inconsistencies and errors.

Another related experiment was carried out by Sormunen (2002), who also repeated the relevance judgment experiment for the same set of queries and results in order to investigate the change in users’ judgments over time. Two rounds of reassessments of TREC-7 and TREC-8 documents with a 4-point relevance scale were performed. However, in the second round of assessments 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, where users were not reminded of their previous assessments, in this study there was a direct and intentional influence of the previous judgments on the changes in the second round’s decisions.

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 record the patterns of these changes in order to analyse how users change their minds over time.

Lewandowski (2008) conducted a user study with 40 subjects who judged relevance (on a binary scale) of the top-20 results from five search engines. He reported quite low precision at 20 results, ranging from 0.37 to 0.52, with Yahoo! and Google outperforming the other search engines and yielding similar results. Vaughan (2004) compared 24 subjects' ranking of four queries' search results with those of Google, AltaVista and Teoma. In her 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 the top-10 results of six search engines on 14 topics and five 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 between 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, of 0.69, to Google 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.

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; in order to find ways 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, while here we investigate the interplay between ranking and relevance judgments Moreover, the gap between rounds in our experiments was one month, and the experimental conditions at each round were identical, apart from the order in which results were displayed to the users.

Method

Experimental setup

Two groups (groups A and B) of graduate (MA) students from the Department of Information Science in Bar-Ilan University were provided with two queries: “Atkins diet” and “Cloud Computing”, and 20 Google search results for each query. For each query, a set of 20 results was constructed: 1) The top-20 Google results for the Atkins diet query (denoted by Google 20); 2) The top-10 Google results and the 101-110 ranked Google results for the Cloud Computing query (denoted by Google 10&100). Each group of students (denoted by A and B) received two queries with the same corresponding set of search results but in different order to neutralize the possible influence of presentation bias on their judgments. In total, in this study, there were four separate relevance judgment experiments: AtkinsA and AtkinsB received the Google 20 result set, while CloudA and CloudB received the Google 10&100 result set.

The four experiments were run three times (i.e. in three rounds) with a one-month interval between each round. The results were presented in a Google form, and in a different random order during each round, to prevent the students from memorising and copying their previous answers. Each student worked independently and had to create between 2-10 categories for the 20 given results per query, and rank these categories according to their degree of relevance to the query. The most relevant category was ranked 1, and the higher the relevance value assigned to the category the less relevant it is to the query. The task given to the students was to gather information in order to write a report/summary of the query topic, although they did not have to submit the summary. The relevance criterion was whether, and to what extent, a result contains information which contributes to the summary of the topic. Group A originally comprised 28 students, and group B comprised 25 students. Several students' judgments were excluded, since they were incomplete, yielding 24 students in 3 experiments (AtkinsA, AtkinsB, CloudB) and 22 students in the 4th experiment (CloudA). The students’ age was between 25-40, 42% of the students were male and 58% female in each of the groups (except for CloudA, where only 36% were male).

Since the queries were predefined and not chosen by the students, we wanted to estimate their interest and knowledge of these topics. Thus, in the experiment form we also added two corresponding questions: “How much are you interested in the presented topics?” and “How do you estimate your knowledge of the topics prior to the experiment?”. The possible answers were on the Likert 1-5 scale, when 5 is the highest level of interest / knowledge and 1 is the lowest. As two open questions, the students were also asked to explain their strategy of grouping results into categories and their method of assessing the relevance of the results' categories. At last, they also had to mention whether or not they changed their method of categorisation in different rounds of the experiments. The replies were coded as a yes or no answer. The students answered the open questions as free text, which was then manually analysed by the authors in order to identify themes which appeared in at least two students' responses.

Measures and quantitative analysis

We computed the average number of categories and the average size of each category (the average number of results in a category) for each experiment and round. To analyse the results, we also calculated the following measures:

1) The average overlap between the results in each category during different round pairs, i.e. the average proportion of the same results in the same category per student in two different rounds of an experiment,
2) The proportion of students who chose the same number of categories for a given query, in each round of the same experiment,
3) The average proportion of categories with the same size, per student in each round, in order to estimate the change in categorisation strategy and relevance judgments of the results over time.

To measure the agreement between Google's search results ranks and the subjects' category ranks, we calculated the average Google rank of the search results in each user-defined category. It has been shown in the literature that there is a substantial gap between Google's and users' rank at the individual result level. Hence, it is important to estimate the agreement between the search engine (Google in our case) rank and users' categorical relevance rank, in order to explore whether the gap between users' and search engine's ranks is still present when taking the relevance categories-based approach and to what extent. To measure the overall agreement, we devised a novel similarity measure, called the *average concordance*, which is calculated by performing the following steps per student:

1) For a pair of categories c1 and c2 such that relevance(c1) > relevance(c2):

   If the average Google rank for category c1 is higher than that of c2 (i.e. as expected there is a match and both student's relevance and average Google's rank are higher for c1 than for c2) – assign 1 to the counter, otherwise the counter remains 0;

2) Repeat step 1 above for all the pairs of categories for a user;
3) Sum up the counters for all the pairs and divide by the number of pairs.

In addition, we present another novel similarity measure, called the *MinMax swap ratio*, in order to calculate more precisely (than the average concordance) the agreement level between Google's ranks and the users' category relevance judgements. This is a pairwise measure, returning a number between 0 and 1, which computes the degree to which a category of higher relevance, c1, agrees with a category, c2, of lower relevance. The algorithm basically computes the proportion of swaps required for a pair of categories to make the more relevant category contain higher ranked Google results than the less relevant category, defined more precisely as follows:

1) For each pair of user constructed categories c1 and c2 such that relevance(c1) > relevance(c2) according to the user's judgment:
    a. Find the result with the top Google rank for c2 (max2) and the result with the lowest Google rank for c1 (min1);
    b. If max2>min1 then swap these results (the result of max2 moves to category c1 and the result of min1 moves to category c2);
    c. Add 1 to the count of swaps;
    d. Repeat steps a-c for c1 and c2 after the initial swap until the condition in step b is false;
    e. Normalise the resulting count of swaps by dividing it by the size of the smaller category out of the two categories in the examined pair.

2) Repeat step 1 for all the pairs of categories for the user;
3) Calculate the average swap ratio for all the pairs of categories and then subtract it from 1 to get a similarity score instead of the distance.

We also qualitatively analysed and summarised the students' replies to the open questions regarding their method of work during the experiments.

Results

Quantitative analysis

Table 1: Descriptive analysis for the four experiments.

	User interest in the topic (on a Likert scale 1-5) – average and standard deviation	**User knowledge of the topic (on a Likert scale 1-5) – average and standard deviation**
AtkinsA	3.63(1.28)	3.88(1.08)
AtkinsB	3.54(1.10)	3.71(0.95)
CloudA	3.41(0.96)	3.68(1.04)
CloudB	2.58(1.02)	3.17(1.17)
Average	3.29(1.09)	3.61(1.06)

As can be seen in Table 1, the users had considerable interest and knowledge of both topics, although less interest and knowledge for Cloud Computing, which was more pronounced in group CloudB. Overall, there is a very high (r=0.989) statistically significant (p<0.01) Pearson correlation between user interest in the topic and their knowledge of the topic for all the queries and groups.

Table 2: Descriptive analysis for the number of categories in the four experiments.

	How many subjects did not change no. of categories in all three rounds	**Min no of categories for r1**	**Min no of categories for r2**	**Min no of categories for r3**	**Max no of categories for r1**	**Max no of categories for r2**	**Max no of categories for r3**	**Average no. of categories for r1**	**Average no. of categories for r2**	**Average no. of categories for r3**
AtkinsA	**33%**	**3**	**2**	**2**	**8**	**8**	**8**	5.17 (1.51)	4.38 (1.44)	4.38 (1.47)
AtkinsB	**21%**	**3**	**2**	**2**	**10**	**6**	**6**	5.12 (1.69)	4.42 (0.99)	4.42 (1.12)
CloudA	**32%**	**2**	**3**	**2**	**9**	**6**	**6**	4.45 (1.50)	3.86 (1.02)	3.86 (0.99)
CloudB	**29%**	**3**	**2**	**2**	**10**	**5**	**6**	4.46 (1.98)	3.88 (0.93)	4.0 (1.24)
Average	**30%**	**2.75**	**2.25**	**2.00**	**9.25**	**6.25**	**6.50**	4.80 (1.67)	4.14 (1.10)	4.17 (1.21)

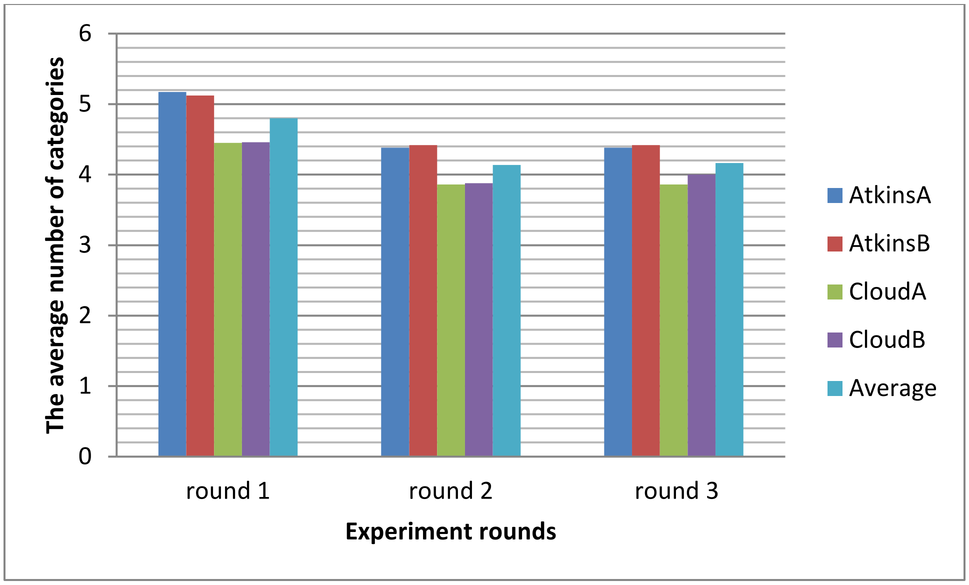


Figure 1: The average number of categories for different experiments and rounds.

As can be seen in Table 2, the majority of the students in both groups created 4-5 categories in each experiment and round. Moreover, as shown in Figure 1 the number of categories is the highest in round 1, decreases for the 2nd round and stabilises in the 3rd round. This result is consistent with the Markov chain model defined in (Zhitomirsky-Geffet, Bar-Ilan, & Levene, 2016a), for determining how users change their minds from round to round until their assessments become stable and converge. Over 80% of the changes between the rounds were the result of adding or removing a single category. There were consistently more categories created for the Atkins diet query by both groups than for the Cloud Computing query, which might be attributed to a wider knowledge and understanding of the former query. Group A chose less categories than Group B for both queries.

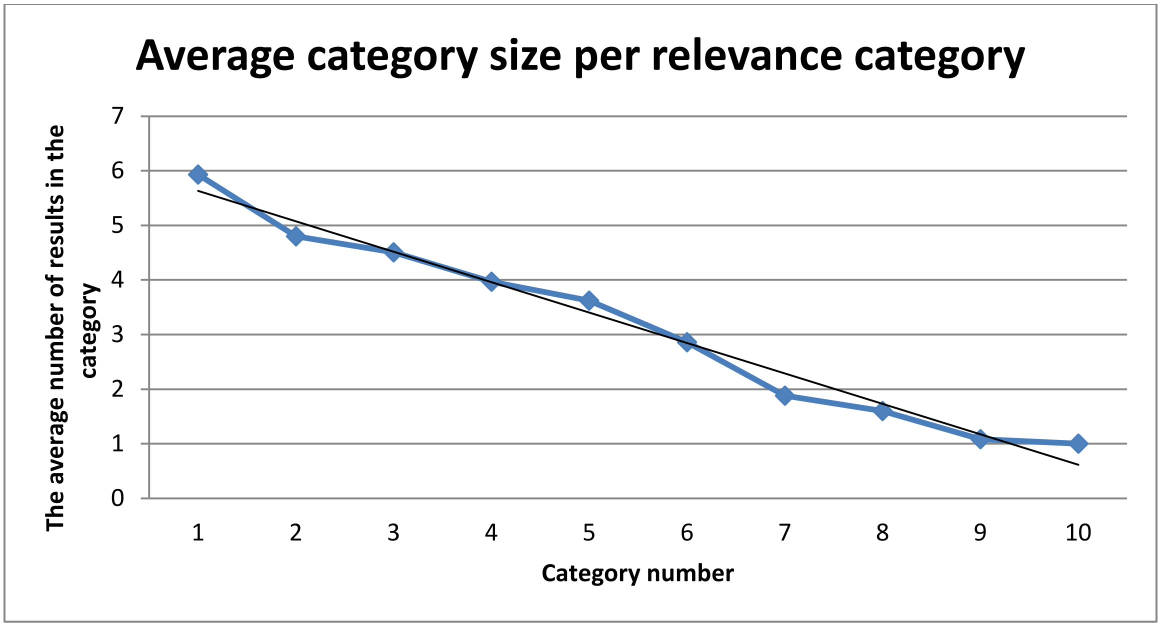
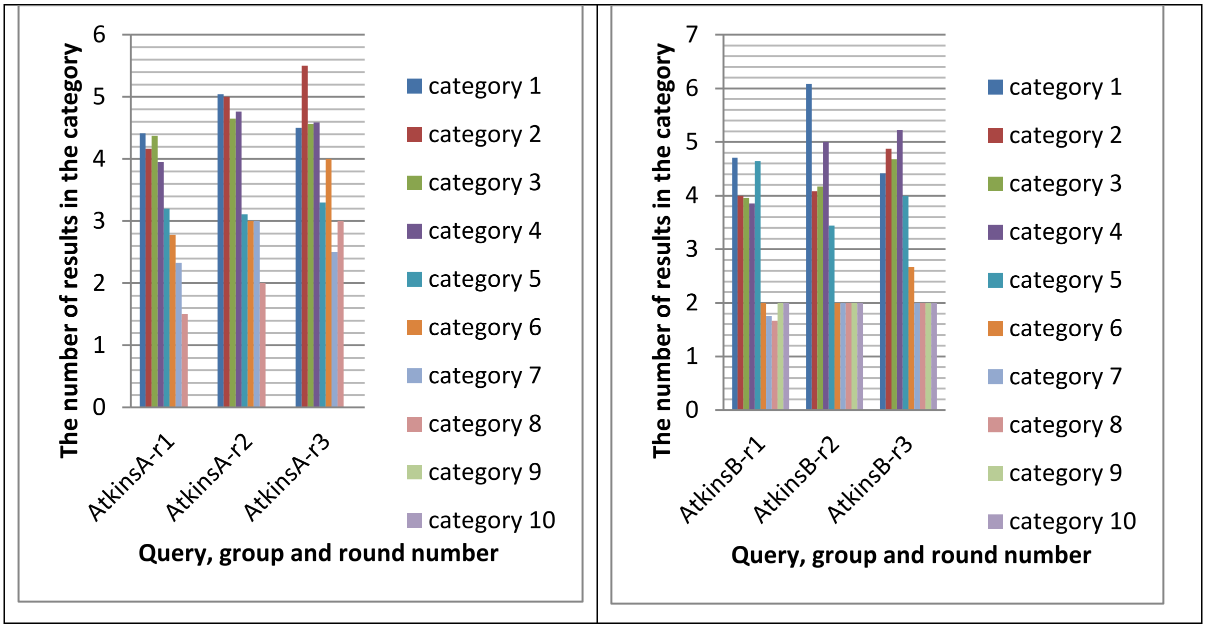


Figure 2a: The average size of the categories for all the experiments.

As can be seen in Figure 2a the category size decreases approximately linearly as the number of category increases (the category becomes less relevant). The R-squared measure for the linear regression is 0.979 and significant at p<0.0001.



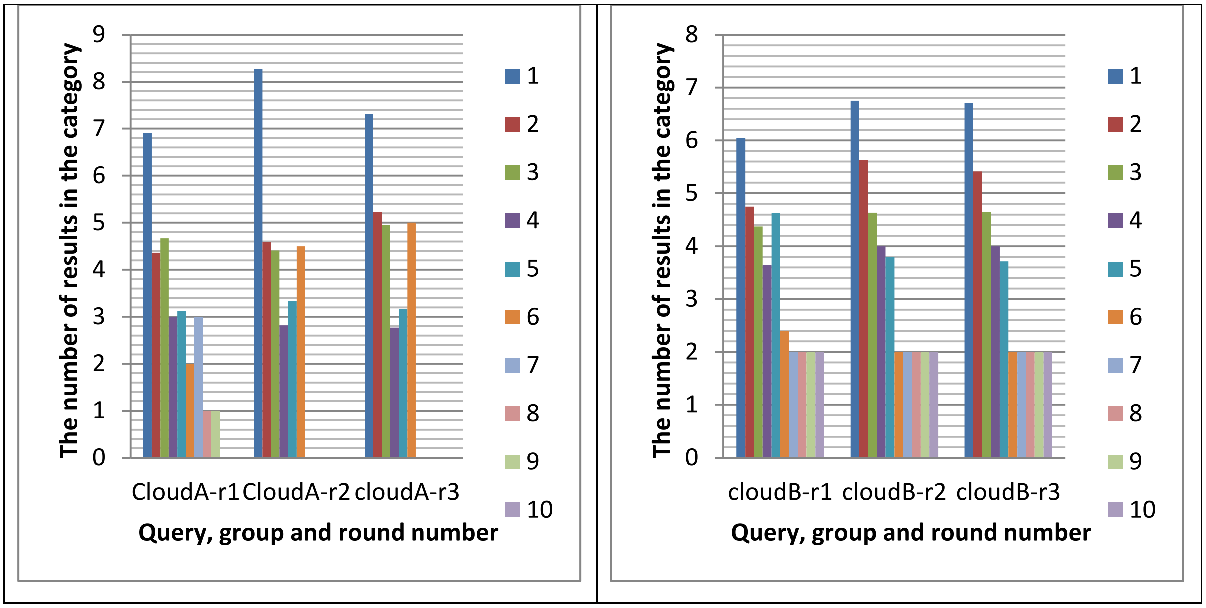


Figure 2b: The size of the categories for different experiments and rounds averaged over all the students (1 – the most relevant category, 10 - the least relevant category; (r1 – round1, r2 – round 2, r3 – round 3).

For the Cloud Computing query, the size of the top (most relevant according to student's judgments) category was always larger than of the other categories as can be seen in Figure 2b. It can be observed from the embedded figures that, for the Cloud Computing query, the gap between the top, most relevant category, and the rest of the categories is much larger than for the Atkins diet query. However, for the Atkins diet query, the distribution of category sizes is, on average, more evenly spread, i.e. the average gap between the sizes of adjacent pairs of categories is 0.42 for Atkins diet query, and it is 0.67 on average for Cloud Computing query. Also, for Atkins diet, in some cases the top category is not even the largest category. In general, the sizes of the categories and the number of the categories were quite stable over time.

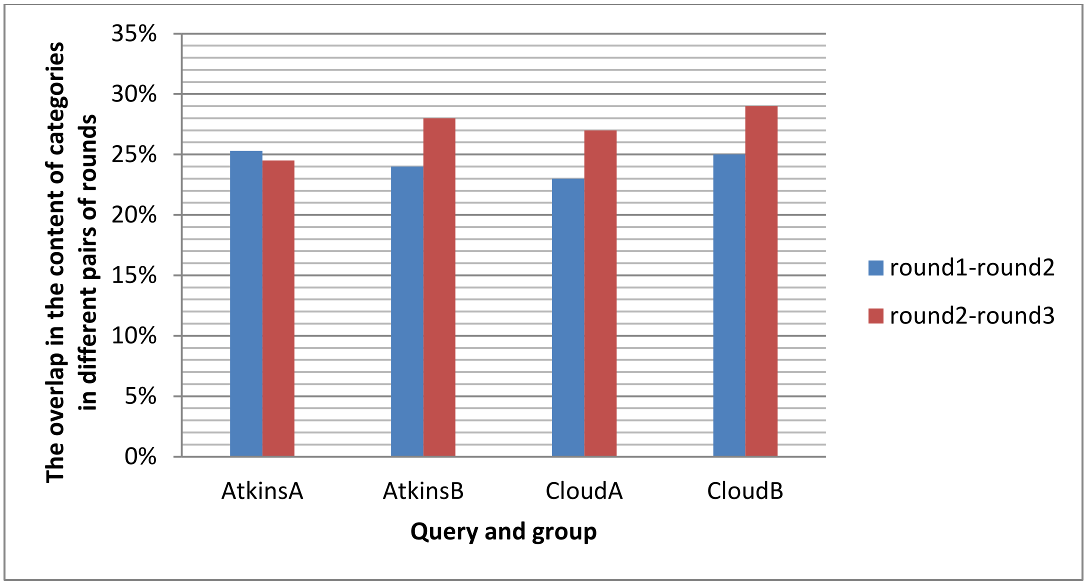


Figure 3: The average level of overlap (in %) between the content of the categories in different rounds (compared for round1 vs. round2 and round 2 vs. round 3) for different experiments.

As can be seen in Figures 3, 4 and 5, the highest overlap in categories' contents, numbers and sizes is usually between rounds 2 and 3, supporting our claim that the assessment tends to stabilise over time.

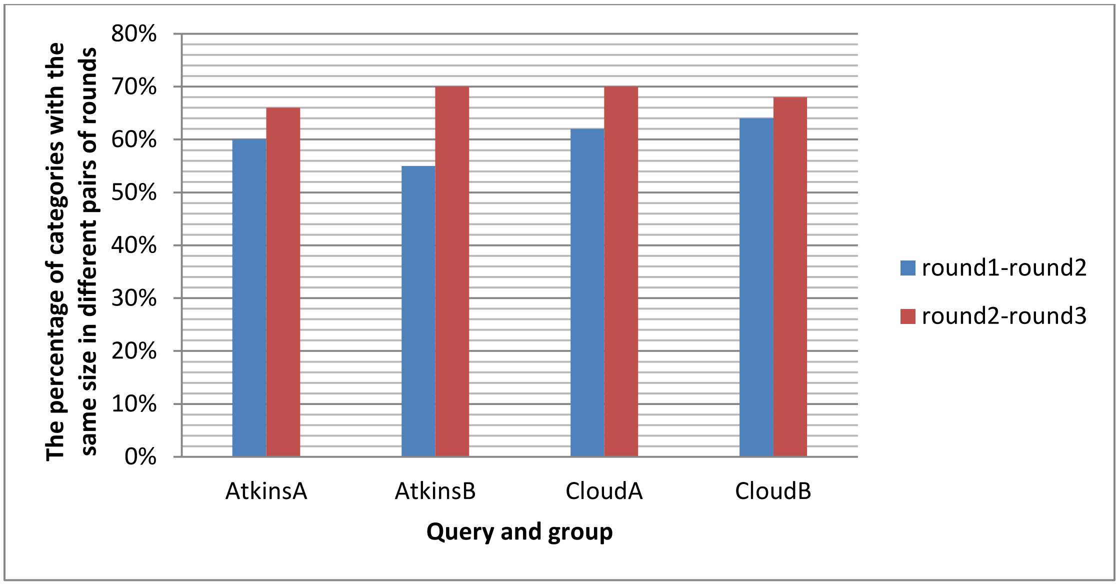


Figure 4: The overlap in category sizes (in %) in different rounds (compared for round1 vs round2 and round 2 vs round 3) for different experiments.

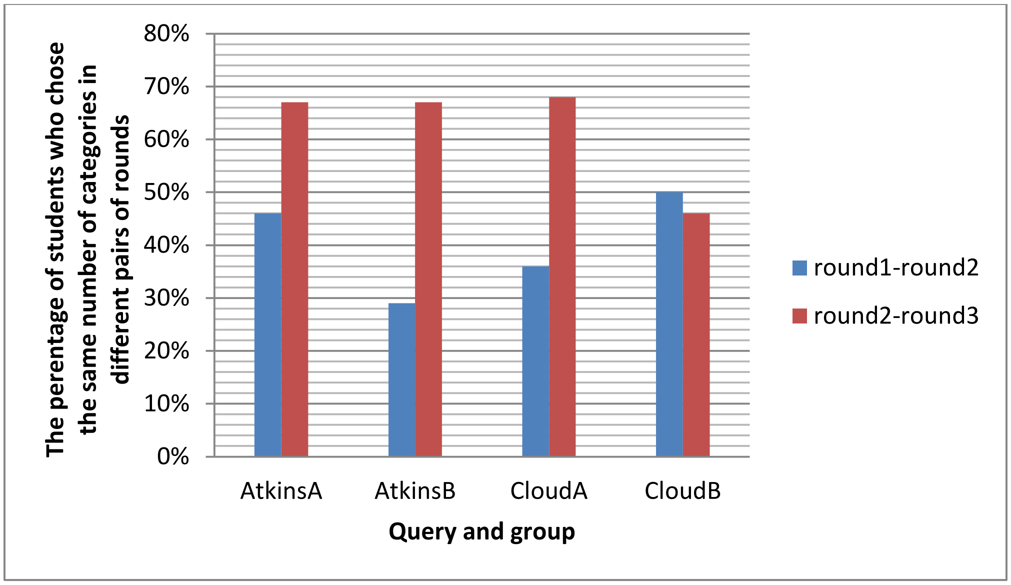


Figure 5: The percentage of students who chose the same number of categories in each pair of rounds (in %) (compared for round1 vs round2 and round 2 vs round 3) for different experiments.

Comparison between subjects' category-based and Google rank-based relevance scores

Here we assess how closely subjects' category-based relevance judgements match the relevance scores implied by Google's ranking using our novel similarity measures – the average concordance and the MinMax swap ratio.

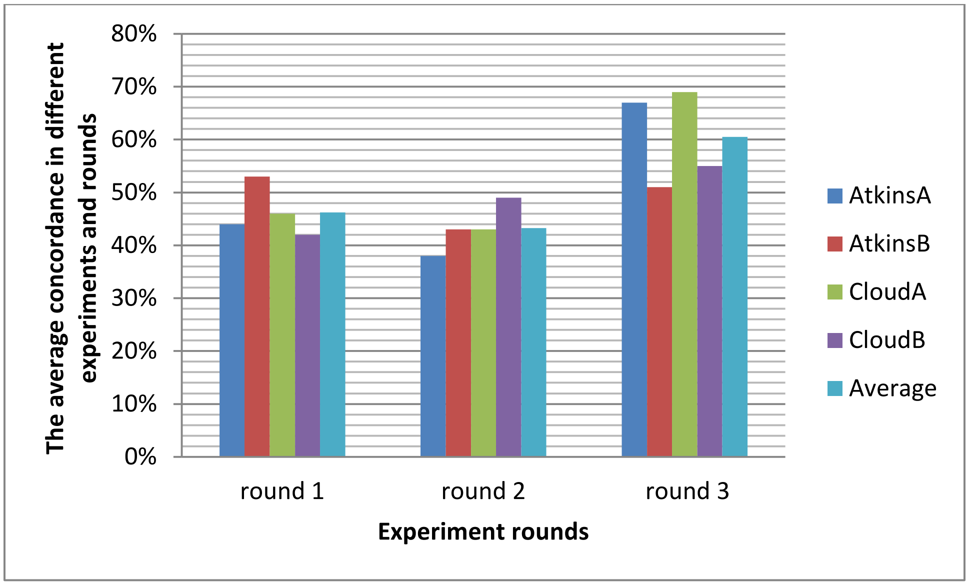


Figure 6: The similarity between Google's ranks and users' relevance judgments as measured by the average concordance for various experiments and rounds.

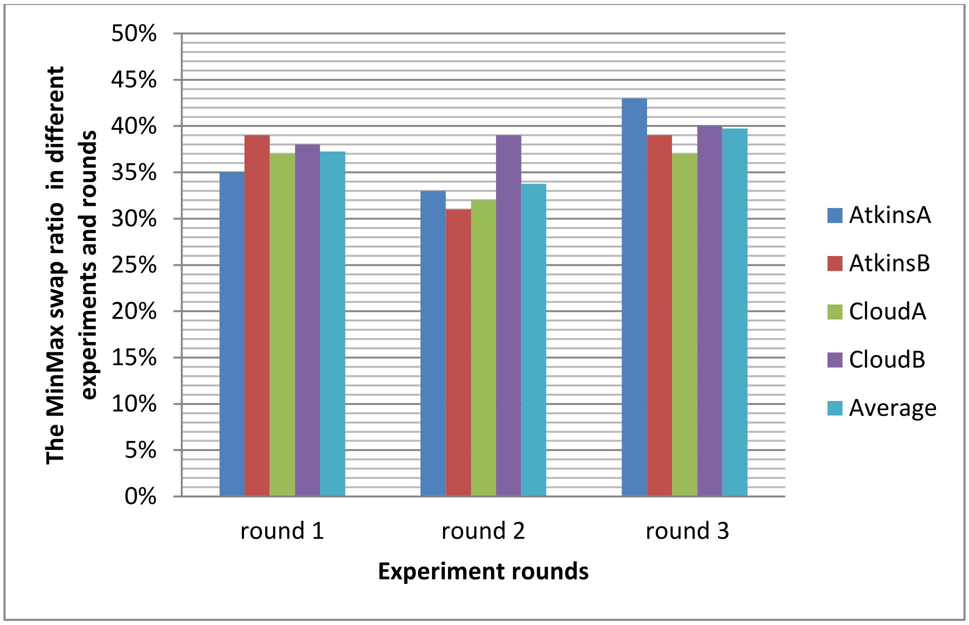


Figure 7: The similarity between Google's ranks and users' relevance judgments as measured by the MinMax swap ratio for various experiments and rounds.

From Figures 6 and 7, it can be observed that the highest agreement according to both measures between the users' category-based relevance and Google's ranking-based relevance, were for the third round of the experiment, virtually for all queries and results sets. This might be explained by the fact that during the first and second rounds the users are learning the topic and the results and more interested in the basic information about it, while in the third round they have already acquired the understanding of the topic as reported in previous work (Serola and Vakkari, 2005), and thus can better assess the quality and relative relevance of the results. We also note that the average concordance leads to a higher similarity than the MinMax swap ratio in all cases, which is due to the fact that the MinMax swap ratio is a finer grained measure.

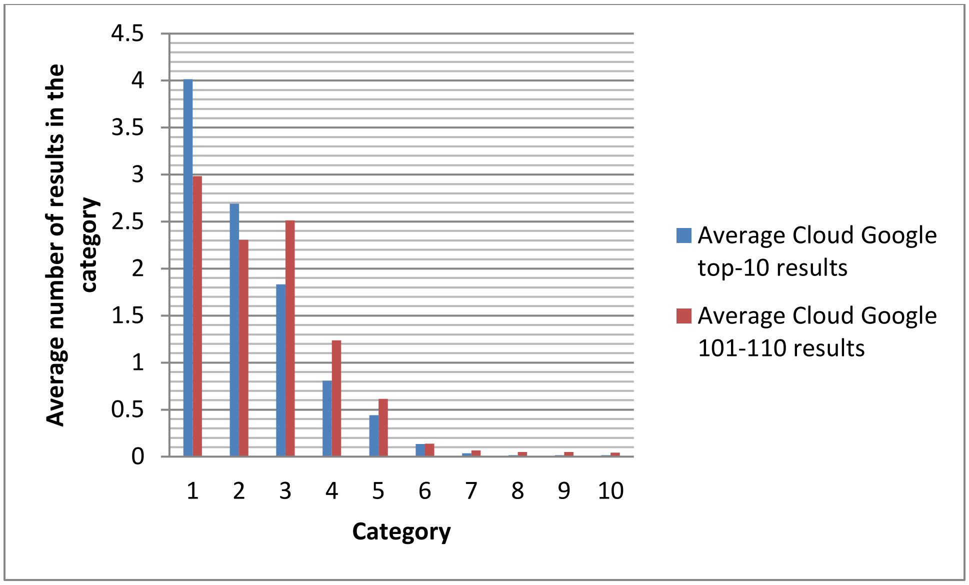


Figure 8: The average number of Google top-10 (1st page) and 10th page (101-110) results in each category for the Cloud Computing query.

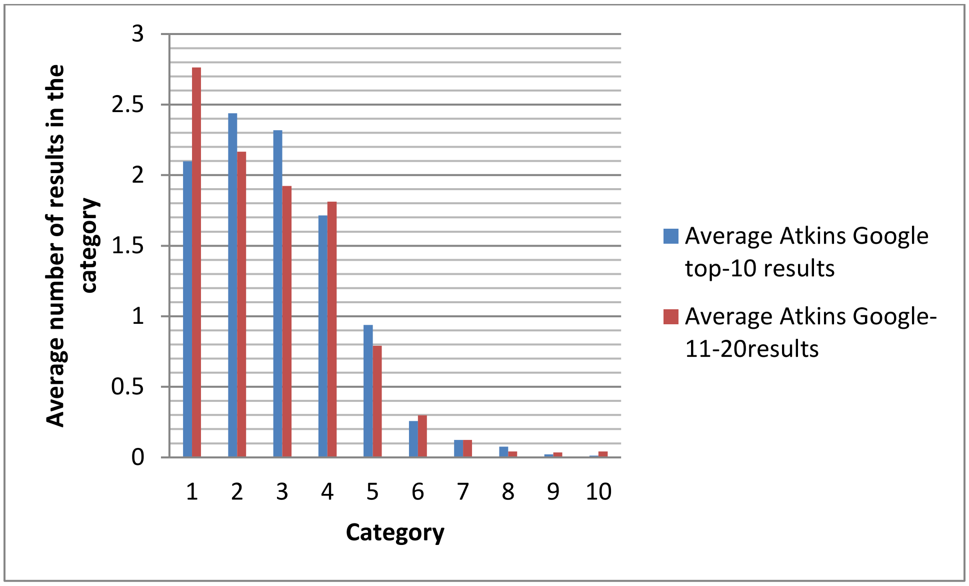


Figure 9: The average number of Google top-10 (1st page) and 2nd page (11-20) results for the Atkins diet query.

We also compared the distribution among the categories of Google's top-10 vs. Google's 11-20 ranked results for the Atkins diet query, and the distribution of Google's top-10 vs. Google's 101-110 ranked results among the categories for the Cloud Computing query. We expected that for the latter query, fewer of Google's 101-110 ranked results will be present in the top-ranked relevance categories than Google’s 11-20 ranked results for the former query. This expectation proved to be correct according to our findings, as shown in Figures 8 and 9 for the first (top-ranked) category. For the Cloud Computing query, the top two categories contained more Google top-10 results than Google 101-110 results, and the next (less relevant) categories contained more Google 101-110 results than Google top-10 results. In contrast, for the Atkins diet query there is no clear pattern, as there were less Google top-10 results in the top category and more Google top-10 results in some less relevant categories (category 4 and 6). A reasonable explanation is that, as one would expect, Google’s 11-20 results are more relevant than its 101-110 results.

Qualitative analysis

In this section we present the main findings from the qualitative analysis of the subjects' replies to the open questions on their working strategy for dividing the results into categories and judging the relevance of these groups to the query. This analysis complements the quantitative analysis of the results and sheds light on the users’ way of thinking when classifying the results into relevance categories. This type of analysis is important, since it uncovers the intrinsic reasons and cognitive processes for categorical relevance judgments, while quantitative analysis shows the results of this type of thinking and processing.

The main findings for the Atkins diet query

The students mentioned the following criteria for assessing the relevance of the results: 1) The level of focus on the topic; 2) the level of coverage (the broader the better); and 3) the level of objectivity, trust and professionalism of the source/author. The results were divided into categories according to their content and source type. For content-based classification according to the level of focus and coverage criteria, the categories’ relevance decreases from specific to more general: 1) Sites directly concentrating on and explaining the main aspects of Atkins diet only, what to eat / how to do this diet in practice; 2) Sites concentrating only on some specified aspects of the Atkins diet, indirectly leading to more pages on Atkins; 3) Pros and cons the Atkins diet, people stories, anti-Atkins opinions, brief mention of Atkins; and 4) Diets in general and alternative diets;

For the source type-based classification strategy, the results were divided into categories according to their objectivity/professional level and level of trust/authority. The category of professional unbiased sites, e.g. medical sites, presenting the advantages and disadvantages of this diet with a lot of references to scientific literature were ranked as the most relevant. This was followed by the category of personal and social sites, such as blogs, forums with people experiences and opinions.;

Finally, the categories of biased ideological sites with agenda, e.g. sites of vegetarian organizations and commercial sites representing products having their commercial interests, were evaluated as the least trustworthy and thus least relevant.

Interestingly, students expressed a negative attitude to Wikipedia results. Wikipedia was considered as insufficient in terms of authority, coverage and professionalism. Also, the result presenting a video on the topic was not, generally, appreciated by the students and was considered a non-direct source.

The main findings for Cloud Computing query

The criteria for assessing the relevance of the results for this query were as follows. Simplicity of the explanation, briefness, examples in the explanation of the concept, coverage and comprehensiveness of the explanation, easiness of access and visualisation of the website, quality/trust/objectivity of the source. The results were divided into categories according to their content and source type, as well as by audience to which they are addressed and their being up-to-date. According to the content type the following categories were composed in decreasing order of relevance: 1) Sites with brief and clear definition and explanation of the concept; 2) Complementary/enriching information beyond the basic definition; and 3) Sites with information mostly non-relevant to the topic. According to their source type and their objectivity/professional level and level of trust/authority the following categories were created in decreasing order of relevance: 1) Wiki and dictionaries, objective sources; 2) Academic sites; 3) News sites, blogs, critical opinions; and 4) Commercial sites of companies and services - biased to their own interests.

From the audience-oriented perspective, the sites comprehensible for laymen/non-professionals/beginners were ranked higher than those addressed to professionals in students' opinion, thus complying with the simplicity and briefness criteria. The sites with most recent and directly related to the topic information were ranked higher than those with less recent or outdated information.

The attitude to Wikipedia results was quite positive for this query. A few subjects used Wikipedia as the baseline for comparison - if the site provides information related to the topic, which is not present in Wikipedia, then it is judged as relevant, otherwise if does not supply any new information compared to Wikipedia - it is considered less relevant. Many students mentioned that the topic was perceived as complicated and new to them, most of them preferred Wikis and dictionaries as the most relevant category of results; several subjects mentioned that all the results were quite closely relevant to the query topic.

Comparative analysis of the findings for the two queries

As can be observed from the findings above, for both queries users divided the results into categories according to their trust and perceived objectivity of the source, and by the specific/direct relatedness of the content to the query. As opposed to the Cloud Computing topic, which was a new topic for most of the users and relatively difficult to learn (according to their feedback), the Atkins diet topic was easier for them to comprehend; even if they were unfamiliar with this specific diet they were quite familiar with the concept of diet in general. In addition, Atkins diet is a controversial topic in its nature, while Cloud Computing is not.

These differences had some influence on the subjects' relevance judgement criteria. For Cloud Computing one of the repeating narratives was choosing the clearest and simplest description/definition of the concept as the most relevant, and consequently putting dictionaries/encyclopaedia/wiki sites into the top category, while for Atkins diet, Wikipedia was ranked lower, and the medical/professional sites with the most broad and comprehensive information representing both positive and negative evidence on the topic were chosen as the most relevant. In other words, simplicity of definition for a layman and short summarisation as criteria for relevance for Cloud Computing versus comprehensiveness and professionalism of the information for Atkins diet. In addition, for Atkins private users' information sources telling their experience/s stories when doing the diet, were ranked quite highly by the subjects (as a second-top category), while this kind of information source was not relevant in judging the relevance of Cloud Computing search results. For Cloud Computing the examples of usage of this technology were highly appreciated by our subjects. For Cloud Computing only, the issues of the source being up-to-date, the sites' visual/GUI attractiveness, and simplicity of explanation, including examples, were considered as important criteria for relevance assessment.

For both queries, users divided the results into groups and then assigned each group a relevance rank. Users wanted first to see the category with results explaining the topic objectively, then they wanted to get more general/complimentary/subjective the opinion information, and there was virtually a complete consensus regarding low ranking of commercial sites and sites with a clear or hidden agenda/bias. For both queries, 3-4 categories were frequently argued to be the most optimal number of categories. For both queries, about 33% reported not changing the method of category choice strategy and relevance judgment in different rounds. The reason for changing the strategy in further rounds was due to learning and gaining more knowledge about the topic from the previous rounds.

In summary, for both queries, the subjects demonstrated a natural ability to divide the results according to their content specification, with direct focus on the topic of the query only, coverage and comprehensiveness, or according to its source type and trustworthiness. Finally, few students tended to combine several criteria in their relevance judgments of the results.

Discussion and Conclusion

In this study we systematically explored the hypothesis of categorical judgment of search results. Our findings show that users typically form 3-5 categories as the optimal number of categories when grouping 20 search engine results. At the beginning of the process they tend to choose more categories, and thereafter the numbers decrease and stabilise during the 2nd and 3rd rounds. The figures for the Atkins diet query results were always higher than for the Cloud Computing results, which might be explained by the controversy surrounding the Atkins diet. In addition, since users' background knowledge and understanding of the search results for the Atkins diet query were better, their categorisation of the results was more fine-grained and more evenly distributed between the categories than for the Cloud Computing query, where most of the results were classed into the top categories.

Moreover, users tend to put more results into the top relevant categories (the average category size for the top category is 5.9) and leave only few results for the less relevant categories (the average category size for the 6th ranked category is only 2.8). For the Cloud Computing query, the top category was larger than for the Atkins diet query in all of the experiments. This might be explained by the fact that, especially for the Cloud Computing query, most of the search results were quite relevant to the query, as was explicitly shown in the users' answers to the open questions, even though half of the results were ranked above 100 by Google. From the analysis of changes over time it seems that for all the parameters the least change was obtained between rounds 2 and 3 thus implying that the assessment process tends to stabilise over time, in agreement with the Markov chain model for how users’ change their minds when assessing search engine results (Zhitomirsky-Geffet, Bar-Ilan, & Levene, 2016a). The changes in users' behaviour over time can be attributed to the learning process they experience during the first rounds (Vakkari, 2016).

From the qualitative analysis of the user comments on the open questions, the grouping of results into categories seemed to be very natural and clear for the users, as well as the assignment of a relevance to each category. We found that the main criteria were direct focus, clear definition and coverage of the specified subject matter of the query, the source being objective and unbiased, and presenting a variety of opinions on the subject of the query, which supports previous findings in (An et al., 2013). Commercial sites were typically considered as unreliable and were categorised as the least relevant. Interestingly, the users' attitude to Wikipedia was different for the two queries. It was ranked higher for the Cloud Computing query, which was perceived as less familiar and more complicated by the users, than for the Atkins diet query. This finding might imply that when users have enough knowledge and understanding of the query topic they prefer more academic and professional information sources, however, when their prior knowledge is limited they prefer Wikipedia and other online encyclopaedic sources.

The significance of this study is that its results provide a proof-of-concept for validity of the categorical thinking hypothesis in search result evaluation. Both quantitative and qualitative analysis support the hypothesis and suggest that users' intrinsically divide the results into categories of similarly relevant results. Furthermore, our work on applying categorical thinking to the grouping of search results into relevance categories according to their content and type of information source, may advise search engines on a sensible categorisation strategy of the top results presented to the user. The outcomes of this research imply that for the top-20 results, 4-5 categories of decreasing size, may help users sift through the information, since each category is "coarse" in the sense that all results in the same category are considered to be of the same relevance to the query. Whether commercial or encyclopaedic knowledge is preferred will, of course, depend on the information goal of the user.

This research also has some limitations. First, this is a user study based on manual evaluation of search results. Second, the study was based on only two queries, due to the fact that asking each user to evaluate more than 40 results for three times (120 results overall) would have been impractical in our case. However, the study provides helpful insights into users' categorical relevance evaluation process, but in order to generalise the conclusions more empirical evidence is required. For instance, as discussed above the sizes of the categories and criteria for their creation may depend on the topic of the query.

References

An, J., Quercia, D., & Crowcroft, J. (2013). Why individuals seek diverse opinions (or why they don't). In: *Proceedings of the 5th ACM Web Science Conference* (pp. 15-18).

Bates, M. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Information Review*, *13*(5), 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*, *58* (9), 1254-1266.

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

Bateman, J. (1998). Changes in relevance criteria: A longitudinal study. *Journal of the American Society for Information Science*, *35*, 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*(9), 1879–1896. doi: 10.1002/asi.22675.

Bruce, H. W. (1994). A cognitive view of the situational dynamism of user centered relevance estimation. *Journal of the Association for Information Science, 45* (5), 142–148.

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.

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

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*(9), 1763–1775. doi: 10.1002/asi.23304.

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with Computers, 10* (3), 303-3202.

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

Ruthven, I., Azzopardi, L., Baillie, M., Bierig, R., Nicol, E., Sweeney, S., & Yakici, M. (2007). Intra-Assessor consistency in question answering. In: *Proceedings of the 30th International ACM SIGIR Conference,* 23-27 July 2007, Amsterdam, Netherlands. (pp. 727-728).

Saracevic, T. (1996). Relevance reconsidered. 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).

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*, *58*(13), 2126-2144.

Scholer, F., Turpin, A., & 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; Beijing, China. New York: ACM, (pp. 1063-1072).

Scholer, F., Kelly, D., Wu, W. C., Lee, H. S., & 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*, New York: ACM, 2013; (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-7: Counting on negligible documents? In: *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, (pp. 324-330)

Spink, A. and Dee, C. (2007). Cognitive shifts related to interactive information retrieval. *Online Information Review*, *31*(6), 845-860.

Smithson, S. (1994). Information retrieval evaluation in practice: A case study approach. *Information Processing and Management*, *30*(2), 205–221.

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*(8), 676–685.

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

Vakkari, P. and Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. *Journal of Documentation*, *56*(5), 540–562.

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 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 the American Society for Information Science*, *32*, 181–188.

Zhitomirsky-Geffet, M. &. Daya, Y. (2015). Mining query subtopics from social tags. *Information Research*, *20*(2), paper 66. Retrieved from http://InformationR.net/ir/20-2/paper666.html.

Zhitomirsky-Geffet, M., Bar-Ilan, J., & Levene, M. (2016a). A Markov chain model for changes in users' assessment of search results. *PLoS ONE* *11*(5): e0155285. doi:10.1371/journal.pone.0155285.

Zhitomirsky-Geffet, M., Bar-Ilan, J., & Levene, M. (2016b). 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), 407-427.

Zhitomirsky-Geffet M., Bar-Ilan J., & Levene M. (2017). Analysis of change in users' assessment of search results over time. *Journal of the Association for Information Science and Technology (JASIST), 68*(5), 1137-1148.