Authors:	Keskustalo Heikki, Järvelin Kalervo, Pirkola Ari, Sharma Tarun, Lykke Nielsen Marianne
Name of article:	Test Collection-Based IR Evaluation Needs Extension Toward Sessions - A Case of Ex-tremely Short Queries
Name of work:	Proceedings of AIRS 2009, the 5th Asia Information Retrieval Symposium
Editors of work:	Lee G & al.
Year of publication:	2009
ISBN:	978-3-642-04768-8
Publisher:	Springer
Pages:	63-74
Series name and number:	Lecture Notes in Computer Science 5839
ISSN:	0302-9743
Discipline:	Natural sciences / Computer and information sciences
Language:	en
School/Other Unit:	School of Information Sciences

URL: http://www.springerlink.com/content/7511872uu5v0q774/fulltext.pdf
URN: http://urn.fi/urn:nbn:uta-3-866
DOI: http://dx.doi.org/10.1007/978-3-642-04769-5_6

Preprint for: Keskustalo, H. & Järvelin, K. & Pirkola, A. & Sharma, T. & Lykke Nielsen, M. (2009). Test Collection-Based IR Evaluation Needs Extension Toward Sessions - A Case of Extremely Short Queries. AIRS 2009, the 5th Asia Information Retrieval Symposium, Sapporo, Japan, October 2009. Heidelberg: Springer, Lecture Notes in Computer Science 5839, pp. 63-74. Full text at http://www.springerlink.com/content/751l872uu5v0q774/fulltext.pdf

Test Collection-Based IR Evaluation Needs Extension Toward Sessions – A Case of Extremely Short Queries

H Keskustalo, K Järvelin, A Pirkola, T Sharma, and [2]M Lykke Nielsen

1University of Tampere, Finland [2]Royal School of LIS, Copenhagen, DK
{heikki.keskustalo, kalervo.jarvelin, ari.pirkola}@uta.fi, tarunbhu@yahoo.co.in,
MLN@db.dk

Abstract. There is overwhelming evidence suggesting that the real users of IR systems often prefer using extremely short queries (one or two individual words) but they try out several queries if needed. Such behavior is fundamentally different from the process modeled in the traditional test collection-based IR evaluation based on using more verbose queries and only one query per topic. In the present paper, we propose an extension to the test collection-based evaluation. We will utilize *sequences* of short queries based on empirically grounded but idealized session strategies. We employ TREC data and have test persons to suggest search words, while simulating sessions based on the idealized strategies for repeatability and control. The experimental results show that, surprisingly, web-like very short queries (including one-word query sequences) typically lead to good enough results even in a TREC type test collection. This finding motivates the observed real user behavior: as few very simple attempts normally lead to good enough results, there is no need to pay more effort. We conclude by discussing the consequences of our finding for IR evaluation.

1 Introduction

Recent studies show that real users of information retrieval (IR) systems search by very short queries but may try out several queries in a session [1-5]. Such queries may consist of only 1-2 search keys. Smith and Kantor [3], and Turpin and Hersh [5] both found that real users successfully compensated for the performance deficiencies of retrieval systems by issuing more queries and/or reading more documents. In real life a searcher typically issues an initial query and inspects some top-N result documents. If no or an insufficient number of relevant documents are recognized, the user may repeatedly launch further queries until the information need is satisfied or (s)he gives up. This setting is different from the Cranfield style IR experiments based on verbose queries and one query per topic. IR evaluation focuses on the quality of the ranked result, measured in terms of available single query metrics, such as mean average precision or cumulated gain or its variants [2, 6]. These metrics do not directly pay attention to query formulation costs, that is, they encourage finding quality at any cost, and short queries are not rewarded for their minor formulation costs.

In the present paper, we take another look at user behavior and IR evaluation. In real IR situations there is a cost associated with initial query formulation and subsequent reformulation. Searchers optimize the total cost-and-benefit of their entire sessions. This may render sessions of short queries reasonable. They allow minimal query formulation costs while taking chances with the quality of results. We call such queries as trivial queries: they employ very few search keys in various combinations. Typical real searchers interact with IR systems using such trivial queries.

We show in this paper that trivial queries surprisingly quickly yield reasonable results. We utilize the TREC 7-8 test collection with 41 topics for which graded relevance assessments are available. We will define idealized trivial query strategies and run systematically constructed sessions seeking to find one relevant document (using two distinct relevance thresholds) in Top-10 of each query, and reformulating the query in case of failure. To render our simulation empirically well-founded, we collected data for query candidates from test persons. Our findings theoretically and experimentally motivate the observed real-life user behavior, which real users must have learned through experience, when interacting with IR systems. As few very simple attempts often lead to good enough results, there is no incentive to pay more effort. In Section 2, we review findings on user behavior and consider the costs and benefits of IR sessions before presenting the research problems. Section 3 discusses the construction of simulated sessions. Experimental results are given in Section 4 and discussed in Section 5. Section 6 presents our conclusions.

2 Session costs and benefits

2.1 User behavior

Real searchers behave individually during search sessions. Their information needs may initially be muddled and change during the search process; they may learn as the session progresses, or switch focus. The initial query formulation may not be optimal and the searchers may need to try out different wordings [2]. In fact, it may be impossible for the searcher to predict how well the query will perform [7] because even if the query describes the topic well, it may be ambiguous [8] and retrieve documents not serving the particular information need. Therefore in real IR it is very common that the users may have to revise their topical queries.

Real-life searchers often prefer very short queries [1, 4]. They may also avoid excessive browsing [1, 9]. Jansen and colleagues [1] analyzed transaction logs containing thousands of queries posed by Internet search service users. They discovered that one in three queries had only *one* term. The average query length was 2.21 terms. Less than 4 % of the queries in Jansen’s study had more than 6 terms. The average number of terms used in a query was even smaller, 1.45, in a study by Stenmark [4] focusing on *intranet* users.

The stopping decisions regarding browsing the retrieved documents depend on the search task and the individual performing the task [2]. Jansen and colleagues [1] observed that most users did not access results beyond the first page, i.e., the top-10 results retrieved. Therefore real life sessions often consist of sequences of very short queries. The data in Table 1 reflect these findings.

The data for Table 1 come from an empirical, interactive study comparing two search systems. Thirty domain experts each completed the same four realistic search tasks A – D simulating a need for specific information required to make a decision in a short time frame of several minutes. Each task formed a separate query session. The data represent the sessions of one of the systems, showing great variability between the tasks along various variables. Essentially, there were 2.5 queries per session and 2.4 unique keys per session. On average, each query had two keys and 0.9 filters (a geographic, document type or other condition). Only 10 among the 60 sessions employed four or more unique search keys. These searchers were precision-oriented, i.e., they quit searching soon after finding one or a few relevant documents. The four bottom lines report the frequency of the query strategies (S1-S3) that we shall define in Section 3.3. The total number of identified strategies (72) exceeds the number of sessions (60) because more than one strategy was employed in some sessions.

Table 1. Real-life session statistics based on 15 sessions for Tasks A-D (N=60) sessions [10].

Variable	A	B	C	D	Tot
Tot # queries per task	25	59	28	40	152
Avg queries in session	1.7	3.9	1.9	2.7	2.5
Avg # keys per session	1.5	3.9	1.9	2.2	2.4
Avg # keys per query	1.4	2.4	1.8	2.0	2.0
Avg # filters per query	1.2	1.1	0.8	0.7	0.9
S1 frequency	11	3	4	3	21
S2 frequency	2	4	3	4	13
S3 frequency	4	13	11	10	38
S1-S3 frequency sum	17	20	18	17	72

Real life searching of the kind described in Table 1 is fundamentally different from the Cranfield type IR evaluation scenario. In the traditional test collection-based evaluation a single query per topic exists and the queries used are longer (typically 7 to 15 search keys, see [1]). Because of these facts we will focus on trivial query sessions in the present study, including one-word queries for each topic.

2.2 Identifying costs and benefits

What explains the great difference between user behavior and effective laboratory queries? We believe that costs and benefits of IR interaction are currently not sufficiently taken into account to explain user behavior.

Early papers on IR evaluation had a comprehensive approach toward evaluation: Cleverdon, and colleagues [11] identified, among others, presentation issues and intellectual and physical user effort as important factors in IR evaluation. Salton [12] identified user effort measures as important components of IR evaluation. More recently, Su [13] compared 20 evaluation measures for interactive IR, including actual cost of search, several utility measures, and worth of search results vs. time expended.

Due to time pressure, documents retrieved in the top ranks may be of interest for real users [9, 14]. Järvelin et al. [2] extended the Discounted Cumulated Gain metric

[6] into a session-based evaluation metric which evaluates multiple query sessions and takes the searcher's effort indirectly into account. Also the literature on usability has a comprehensive approach to costs and benefits, see, e.g., the ISO standard [15].

One may conclude that various costs and benefits of interactive IR systems have been brought up in the literature. The same does not hold on current IR evaluation practices. In interactive settings both costs and benefits are present and affect searcher behavior (e.g., through expectations). Therefore, interactive IR evaluation should incorporate the existing cost factors: search key generation cost, query execution cost, result scan cost, next result page access cost, and relevant document gain. Contemporary IR evaluation effectively assumes all costs as zero, thus focusing on benefits (the gain) at any cost. This hardly models real-life situations.

In the present paper we acknowledge that the query formulation costs may be a significant factor explaining user behavior. We will show that trivial queries are a reasonable alternative for the user because their formulation costs are minimal and their effectiveness competitive if sessions are allowed.

2.3 Research problem

Our background assumption is that the observed user behavior [1, 4, 10] does satisfy real needs. Thus, the obvious question is whether it makes sense for the user to combine the use of short queries and generally "take their chances" with trivial queries and reformulate in case of a failure. Our overall research question therefore is: What is the effect of utilizing a sequence of trivial queries as a session compared to the traditional approach of utilizing one verbose query?

In studying this problem, we make simplifying assumptions. First, we assume that the topical requests remain unchanged during a session: the simulated searcher neither learns nor switches focus during the session. Secondly, the simulated searcher is able to recognize the relevance of documents (see [16]). Third, the simulated searcher is assumed to scan the ranked list of documents from the top to bottom (see [17]). However, reflecting the observed searcher behavior, we focus on the Top-10 results. We will focus on the search task of finding a single relevant document (see [18]).

3 Constructing simulated sessions

We made use of the TREC 7-8 test collection, and real test persons to generate candidate queries and alternative search keys. The collection of these data is explained next and their properties analyzed thereafter, followed by the definition of the simulated session strategies, the retrieval protocol, and our evaluation method.

3.1 The test collection and search engine

We used the reassessed TREC test collection including 41 topics from TREC 7 and 8 ad hoc tracks [19]. The document database contains 528 155 documents organized under the retrieval system *Lemur*. The database index is constructed by lemmatizing the document words. The relevance judgments were based on topicality using a four-point scale: (0) irrelevant document: the document does not contain any information about the topic; (1) marginally relevant document: the document only points to the topic but does not contain more or other information than the topic description; (2) fairly relevant document: the document contains more information than the topic description but the presentation is not exhaustive; and (3) highly relevant document: the document discusses the themes of the topic exhaustively. In the recall base there are on the average 29 marginally relevant, 20 fairly relevant and 10 highly relevant documents for each topic [19].

3.2 Collecting the query data

As session-based collections do not currently exist we decided to construct one by ourselves on top of the TREC 7-8 test collection. 41 topics were analyzed intellectually by test persons to form query candidate sets. During the analysis the test persons did not interact with a real system. They probably would have been able to make higher quality queries had they had a chance to utilize system feedback.

A group of seven undergraduate information science students (Group A) and seven staff members (Group B) performed the analysis. Staff members having an extensive background regarding the specific test collection were excluded. Regarding each topic a printed topic description and a task questionnaire were presented for the test persons. Each of the 41 topics was analyzed twice - once by a student and once by a staff member. The users were asked to directly select and to think up good search words from topical descriptions and to create various query candidates.

First a two-page protocol explaining the task was presented by one of the researchers. Information in the description and narrative fields of the test collection topics were presented for the users. Descriptions regarding non-relevance of the documents were also omitted to make the task more manageable within the time limitation of 5 minutes per topic. The test persons were asked to mark up all potential search words directly from the topic description and to express the topic freely by their own words. Third, they were asked to form various query candidates (using freely any kinds of words) as unstructured word lists: (i) the query they would use first (“1st query”); (ii) the one they would try next, assuming that the first attempt would not have given a satisfactory result (“2nd query”). Finally, the test persons were asked to form query versions of various lengths: (iii) one word (1w), (iv) two words (2w), and (v) three or more words (3w+). The very last task was to estimate how appropriate each query candidate was using a four-point scale.

3.3 Simulated session strategies

Using the query data collected from the test persons we created four simulated session strategies (S1-S4) for the experiments. The strategies S1-S3 model *five-query sessions* (short queries), while Strategy S4 acts as a comparison baseline and utilizes *only one*, long query. Sessions longer than five queries are also relatively rare in real life, see

[6] and Table 1. Data collected from the test persons were used in session strategies S1 to S3. In each session, the simulated searcher inspected at most 50 documents: in S1-S3 at most five distinct Top-10 results, and in S4 at most the single Top-50.

Session Strategy S1: One-word queries only

In S1 strategy we experiment *solely with one-word queries*. Unique individual words are selected randomly from various query types in the following order: “1st query”, “2nd query”; 1w, 2w, and 3w+ queries until five distinct words are collected. Within a session, if any given one-word query does not retrieve a (highly) relevant document within its top-10 ranks, we immediately try out the next one-word query. S1 is justified because (1) extremely short queries dominate in real life, and (2) the strategy was employed 21 times in the 60 real-live sessions of Table 1. Random selection obviously creates some bad one-word queries. We purposefully experimented with such a strategy to explore the effects of allowing bad queries within session - as may very often happen in real life.

Session Strategy S2: Incremental query extension

In S2 strategy we experiment with using incrementally longer queries in sessions. As stated above, our test persons were requested to form one-word (1w), two-word (2w), and longer (3w+) query versions, and the queries they would try first (1st query) and second (2nd query). Here we selected words *left to right* (i.e., not randomly) from each query version (using query versions in the order explained above) until a sequence of five, if possible, unique words w1,…, w5 is formed. These words are used to construct queries of varying lengths (i.e., w1; w1 w2; …; w1 w2 w3 w4 w5) (from 1 to 5 words) for each topic. Within each topical session, the searcher starts with the one-word query. If a query does not retrieve the required (highly) relevant document, the next incrementally longer query is launched. S2 is justified because this strategy was employed in 13 times of the 60 real-live sessions of Table 1. It simulates a lazy searcher who tries to cope with minimal effort and adds one word at a time.

Session Strategy S3: Variations on a theme of two words

In S3 *two core search keys* are fixed to represent the information need and several different third words are tried as variations. S3 is justified as this strategy was employed in 38 of the 60 real-live sessions of Table 1. According to Jansen et al. [1] modifications to successive queries are done in small increments by modifying, adding or deleting keys. We used first three words of the 3w+ query as the starting point, and varied randomly the third word by replacing it with distinct words selected from the 3w+, or from 1st, 2nd, 1w or 2w queries (in that order) if 3w+ ran out of words.

Session Strategy S4: Single verbose query

Session strategy S4 consists of a single verbose query. It contains all the words of the *description* and the *title* field (on the average 16.9 words). Thus, it represents traditional laboratory testing and serves as a baseline.

3.4 Retrieval protocol and evaluation

The run procedure went as follows:

1. Based on session strategies S1 to S4 query sequences were constructed.
2. Top-10 documents were retrieved for each query in S1-S3 and top-50 for S4.
3. Success of each session strategy S1 to S4 was determined.

Stopping decisions often depend on the task, context, personality, and the retrieval results [10]. In this study, the stopping condition was defined as finding one relevant document. Failure was defined as inability to find a relevant document in a session.

Table 2. Effectiveness of session strategies S1 to S4 (User Group A and B) for 41 topics. Legend: number-in-cell denotes the ordinal of the first successful query in finding a relevant document within its top-10 ranks. For session S4 see text below. Hyphen denotes a failure to find a relevant document. Table on the left: *liberal* relevance threshold is used. Table on the right: *stringent* relevance threshold is used.

	Liberal Relevance							Stringent Relevance						
	S1		S2		S3		S4	S1		S2		S3		S4
Topic#	A	B	A	B	A	B	_	A	B	A	B	A	B	_
351	1	2	5	1	3	1	1	1	2	5	1	3	1	1
353	1	1	1	1	1	1	1	-	-	2	2	-	1	1
355	1	2	1	1	1	1	1	1	2	1	1	1	1	1
358	1	1	1	1	2	1	1	1	1	1	1	2	1	1
360	1	2	1	1	1	1	1	2	2	3	3	2	1	1
362	1	1	1	1	1	1	1	1	1	1	1	-	3	1
364	1	1	1	1	1	1	1	1	1	1	1	1	1	1
365	3	1	1	2	1	1	1	3	1	1	2	1	1	1
372	5	2	1	1	1	1	1	-	-	2	2	1	2	1
373	1	1	1	1	1	1	1	1	1	1	1	1	1	1
377	2	-	1	-	1	-	1	2	-	1	-	1	-	1
378	-	-	3	3	1	1	1	-	-	-	-	-	-	-
384	2	1	1	1	1	1	1	-	-	4	3	2	1	2
385	-	-	2	2	1	1	1	-	-	2	2	2	1	1
387	1	1	1	1	1	2	1	2	1	2	2	1	2	1
388	2	3	4	3	1	1	1	-	-	-	-	-	-	4
392	2	1	1	1	1	1	1	2	1	1	1	3	1	1
393	2	1	1	1	1	1	1	2	1	1	1	1	1	3
396	1	3	2	1	1	1	1	1	3	2	1	1	1	1
399	4	-	2	1	1	-	1	-	-	4	2	1	-	2
400	4	2	2	1	1	1	1	4	2	2	1	2	1	1
402	1	1	2	1	1	1	1	-	1	2	1	-	1	1
403	1	1	1	1	1	1	1	1	1	1	1	1	1	1
405	1	2	3	1	1	2	1	-	-	-	-	3	-	2
407	1	1	1	1	1	1	1	2	1	1	1	1	1	1
408	2	1	1	1	1	1	1	-	-	2	3	2	1	1
410	3	2	1	1	1	1	1	3	2	1	1	1	1	1
414	-	-	3	2	1	1	1	-	-	-	-	-	-	-
415	3	1	2	1	1	1	1	-	-	2	5	1	1	1
416	3	1	1	1	1	1	1	3	1	1	1	1	1	1
418	2	2	1	1	1	1	1	2	2	1	1	1	1	1
420	1	1	1	2	1	1	1	1	1	1	2	1	1	1
421	1	2	1	3	1	1	1	-	-	2	3	1	1	1
427	2	2	1	1	2	1	1	2	4	-	-	-	-	4
428	2	1	1	1	1	1	1	3	1	1	1	1	1	1
431	2	3	1	1	1	1	1	2	1	1	1	1	1	1
437	-	-	2	3	2	-	3	-	-	-	-	-	-	-
440	-	-	2	2	3	2	1	-	-	2	2	5	-	5
442	1	-	1	2	1	1	1	-	-	-	-	-	-	-
445	3	-	1	1	1	3	1	-	-	2	4	1	4	1
448	-	-	2	2	1	1	1	-	-	-	-	-	-	-

4 Experimental results

Above, general result for session strategies S1-S4 is presented for 41 individual queries (Table 2). Number 1 denotes that the first query in the session was successful in finding a relevant document for the topic within its top-10 ranks. Number 2 denotes the second query being successful etc. The table shows the effectiveness of session strategies based on both *liberal* and *stringent* relevance threshold. The columns for the two searcher groups A and B indicate the variability under the same strategy. Color-coding is used in cells in addition to the ordinal numbers for visual evaluation: black indicates the first query being successful, white no success at all, and grey scale success by a non-first query.

Table 3. Overall effectiveness of session strategies S1 to S4 (User Groups A and B). Top to bottom: average number of queries attempted per session; the count of successful sessions; and the percent of successful sessions.

Liberal Relevance

S1		S2		S3		S4
A	**B**	**A**	**B**	**A**	**B**	-
2.3	2.4	1.5	1.5	1.2	1.4	1.0
35	31	41	40	41	38	41
85.4	75.6	100	97.6	100	92.7	100

Stringent Relevance

S1		S2		S3		S4
A	**B**	**A**	**B**	**A**	**B**	-
3.1	2.9	2.2	2.2	2.2	2.0	1.6
23	23	33	32	31	30	36
60.5	60.5	86.8	84.2	81.6	78.9	94.7

Table 4. Pairwise statistical significance (+) of differences by Friedman's test for Searcher Group A using S1 – S3 and S4, p=0.01.

Liberal Relevance

	Strategies		
	S2-A	**S3-A**	**S4**
S1-A	+	+	+
S2-A		-	+
S3-A			-

Stringent Relevance

	Strategies		
	S2-A	**S3-A**	**S4**
S1-A	+	+	+
S2-A		-	+
S3-A			-

Based on liberal relevance criteria, S1 (*one-word queries only*) is 15-25 percent units weaker than strategies S2-S4 in its success rate (Table 3). Strategies S2-S4 are equally good. The average number of queries in S1-S3 varies between 1.2 and 2.4 queries. Strategy S3 fairs almost as well as S4. On the stringent level S1 is clearly weaker than the other strategies. Surprisingly, S2 and S3 are only 10 – 15 percentage units weaker that S4. The average number of queries in S1-S3 varies between 2.0 and 3.1; the average number of pages browsed for S4 is 1.6.

According to Friedman's test the differences between the strategies are highly significant ($p < 0.001$). Table 4 gives *pairwise* results for Friedman's test and Searcher Group A. We observe that S1 is significantly different from others; S3 is not significantly different from S4, the baseline. The significance results for the Searcher Group B and S4 are similar. Note that even when the results for some trivial query strategies are significantly worse than S4, the queries require much less effort.

Liberal relevance threshold

At the liberal relevance threshold the most successful strategy was the baseline session strategy S4 (*single verbose query*). As only one query candidate was formed in S4 strategy, number 1 in column S4 denotes the fact that the relevant document was found within the first results page (ranks 1 to 10); number 2 denotes second page, etc. All words of the title and description fields were used, thus rendering very long queries - an average query length of 16.9 words per query.

Among the trivial strategies S3 (*variations on a theme of two words*) was the most successful one. For 36 and 34 topics out of 41 (user group A and B, respectively) the very first query was successful. The strategy only failed three times (and only for group B). The session strategy S2 (*incremental query extension*) was also effective. For 27 and 29 topics (for A and B) the very first query (at this point a single key) was successful. Adding the second key to the query helped to find a relevant document for 9 and 7 additional topics (A and B), and the third key for 3 and 4 topics (A and B). Strategy S2 only failed once (in group B). Strategy S1 (*one-word queries only*) was the least successful. The fact that the single query keys were selected randomly obviously hurt the performance. Yet, it failed only in 6 and 10 topics (A and B) out of 41.

Stringent relevance threshold

If liberal relevance threshold is used, low quality documents are accepted as relevant. These documents may be only marginally relevant and escape the reader's attention [16]. Finding one such document can hardly be justified as success even if the user only had to pay minimum effort. Therefore, in the remainder of this paper we only accept highly relevant documents as relevant. They discuss the themes of the topic extensively [19] thus better justifying the user's stopping decision in simulations.

Table 5 summarizes the results based on stringent relevance threshold. The recall base of the test collection did not contain highly relevant documents for 3 topics (#378, #414, #437), leaving 38 topics in the stringent relevance threshold case. Cumulative percentages are shown in the table regarding the share of successful topics for each session strategy.

Also when highly relevant documents are demanded, the baseline session strategy S4 (*single verbose query*) performs best. In 29 topics out of 38 a highly relevant document is found within the first page. Table 5 presents these success figures as percentages (29/38 = 76.3 %). For three topics the second page needs to be inspected in strategy S4; for one topic the third page; for two topics the fourth, and for one topic the fifth page. This strategy fails only for two topics.

Session strategy S3 (*variations on a theme of two words*) was the most successful one among strategies S1-S3 also when highly relevant documents are requested. For 21 and 26 topics out of 38 (for user groups A and B, respectively) the very first query was successful. For 7 and 8 topics (groups A and B) the strategy failed.

Table 5. Success of the session strategies by the ordinal of the query candidate. Figures express the share of the topics (cumulative %) for which a highly relevant document was found. *For explanation regarding session strategy S4 see text.

Query #	S1		S2		S3		S4*
	A	B	A	B	A	B	-
1st	23.7	39.5	44.7	47.4	55.3	68.4	76.3
2nd	47.4	55.3	76.3	68.4	71.1	73.7	84.2
3rd	57.9	57.9	78.9	78.9	78.9	76.3	86.8
4th	60.5	60.5	84.2	81.6	78.9	78.9	92.1
5th	60.5	60.5	86.8	84.2	81.6	78.9	94.7

Table 6. Strategy cost features for group A: strategy (S1-S4), expected number of search keys to enter (T) and queries to launch (Q) for one relevant document.

Strategy	T	Q
S1	8.6	3.5
S2	4.3	2.3
S3	7.3	2.4
S4	16.9	1.0

The session strategy S2 (*incremental query extension*) was also effective. For 17 and 18 topics (A and B) the very first query (i.e., a singe key) attempted was successful. Adding the second key to the query helped to find a highly relevant document for 12 and 8 of the so far unsuccessful topics (A and B). Adding a third key helped to find a highly relevant document for 1 and 3 novel topics (A and B). Only in 5 cases for group A and in 6 cases for group B the incremental extension strategy failed.

The session strategy S1 (*one-word queries only*) was the least successful, yet remarkably, in more than half of the topics (57.9 %, Table 5, groups A and B) the strategy was successful after only three single-word queries were attempted.

We experimented also by measuring the effectiveness of S1 using traditional metrics (P@10 and non-interpolated average precision (AP)). We evaluated the effectiveness of all query candidate sets, 1st to 5th queries for group A and B), averaged over 38 topics, based on the top-1000 documents, and stringent relevance threshold. The highest value observed for P@10 was 7.4 % and for AP 11.3 %. The corresponding values for S4 (verbose queries) were 25.2 % and 19.5 %. Thus, if one query per topic is assumed, S1 queries are inferior, but. sense if multiple-query sessions are used.

5 Discussion

Real users of IR systems typically search by very short queries (often 1-2 keys) but try out several queries if needed [1, 4]. Test collection-based evaluation, e.g., like the one performed in TREC [20] typically employs longer queries and one query per topic. We wanted to analyze the effectiveness of sessions of very short queries. We performed a simulation because it is very difficult to use real interactive sessions and have control over multiple query/session types, avoid learning effects, and support repeatability of the experiment. The strengths of our approach include session strategies and intellectual word selections for queries based on an empirical ground truth.

We therefore had two sets of test persons to create realistic content for trivial queries. Our test persons did not interact with the retrieval system / test collection and thus did not use their own relevance assessments. This is justified as our aim is to study idealized strategies. As we needed to guarantee, that no user learning takes place, our queries may have had lower quality than those in realistic situations. However, if our findings are conservative, this makes the argument only stronger.

The idealized session strategies were constructed based on empirical data (Table 1). We set the limit of maximum of 5 queries per session because longer sessions are rare in real life. We set the limit of maximum of 10 documents per query results for strategies S1-S3 because scanning length in real life is limited [1]. The simplest strategy S1 (*one word queries*) was popular in our sample data (Table 1) and attempts to minimize the query formulation costs. The strategy S2 (*incremental query extension*) was less popular but nevertheless present in the sample data. The strategy S3 (*variations of three word queries*) seeks to fix a focus by two keys and vary by trying out different third keys. This was the most popular strategy in our sample data and to some degree corresponds to Bates' Berry-picking strategy [21]. S4 represents a long TREC-type of single query strategy, and did not occur in our data.

Traditional IR evaluation focuses on the quality of the ranked result. We argue that a more faithful to real-life evaluation should include additional factors: search term generation cost, query execution cost, result scan cost, and next page access cost.

Table 6 shows the expected number of search keys and queries, when each strategy is successful. Regarding S1-S3, we assume that for unsuccessful topics the searcher would in desperation launch one more query (#6), a successful one represented by S4 containing on average 16.9 search keys, to guarantee comparable performance. Strategies S1-S3 yield a low query formulation cost in the number of search terms. If the query launching costs and result scanning unit costs are minor, strategies S1-S3 make sense to users. They mean low formulation costs while taking chances with the result. Our simulations suggest that trivial query sessions make sense. We focused on the limited task of finding one relevant document but our simulation method fits well to tasks where more than one relevant document is required.

6 Conclusion

Log analyses reveal that real users often try out sessions of several short queries. Traditional laboratory evaluation is not well-suited to study this phenomenon.

In this paper we demonstrated session-based batch evaluations utilizing test collections, and query data collected from test persons. We focused on studying the effectiveness of *sessions of very short queries*. We assumed searchers requiring only one relevant document, who browse a very limited length of results, and use a limited set of session strategies (S1-S4). Short query sessions turned out to be successful. For example, our most extreme pure one-word strategy (S1) found the required highly relevant document for most topics if just three words were tried out.

Future test collection-based IR evaluation should model (i) processes where the searcher may try out several queries for one topic and (ii) broader costs and benefits than the ones focusing on the quality of the retrieved result. We believe that including these viewpoints in evaluation may help toward resolving the current disparity of the observed searcher behavior and the assumptions of laboratory experiments.

References

1. Jansen, M. B. J., Spink, A., Saracevic, T.: Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web, Information Processing & Management, 36(2), 207-227 (2000)
2. Järvelin, K., Price, S. L., Delcambre, L. M. L., Nielsen, M. L.: Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions, in ECIR'08 (2008)
3. Smith, C. L., Kantor, P. B.: User Adaptation: Good Results from Poor Systems, in Proc. ACM SIGIR'08, 147-154 (2008)
4. Stenmark, D.: Identifying Clusters of User Behavior in Intranet Search Engine Log Files. Journal of the American Society for Information Science and Technology, 59(14), 2232-2243 (2008)
5. Turpin, A., Hersh, W.: Why Batch and User Evaluations Do Not Give the Same Results, in Proc. ACM SIGIR'01, 225-231 (2001)
6. Järvelin, K., Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques. ACM TOIS, 20 (4), 422-446 (2002)
7. Swanson, D.: Information Retrieval as a Trial-and-Error Process. Library Quarterly, 47(2), pp. 128-148 (1977)
8. Sanderson, M.: Ambiguous Queries: Test Collections Need More Sense, in Proc. ACM SIGIR'08, 499-506 (2008)
9. Azzopardi, L.: Position Paper: Towards Evaluating the User Experience of Interactive Information Access Systems, in SIGIR'07 Web Information-Seeking and Interaction Workshop, 5 p. (2007)
10. Lykke, M., Price, S., Delcambre, L., Vedsted, P.: How doctors search: a study of family practitioners' query behaviour and the impact on search results (In press) (2009)
11. Cleverdon, C. W., Mills, L., Keen, M.: Factors determining the performance of indexing systems, vol. 1 - design. Cranfield: Aslib Cranfield Research Project (1966)
12. Salton, G.: Evaluation Problems in Interactive Information Retrieval. Information Storage & Retrieval, 6, 29-44 (1970)
13. Su, L. T.: Evaluation Measures for Interactive Information Retrieval. Information Processing & Management, 28(4), 503-516 (1992)
14. Hersh, W.: Relevance and Retrieval Evaluation: Perspectives from Medicine. Journal of the American Society for Information Science, April 1994, 201-206 (1994)
15. ISO: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs), Part 11: Guidance on Usability. ISO 9241-11:1998 (E) (1998)
16. Vakkari, P., Sormunen, E.: The Influence of Relevance Levels on the Effectiveness of Interactive Retrieval. Journal of the American Society for Information Science and Technology, 55(11), 963-969 (2004)
17. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately Interpreting Clickthrough Data as Implicit Feedback, in Proc. ACM SIGIR'05, 154-161 (2005)
18. Price, S. L., Nielsen, M. L., Delcambre, L. M. L, Vedsted, P.: Semantic Components Enhance Retrieval of Domain-specific Documents. In Proc. CIKM'07, 429-438 (2007)
19. Sormunen, E.: Liberal Relevance Criteria of TREC - Counting on Negligible Documents? In Proc. ACM SIGIR '02, 324-330 (2002)
20. Voorhees, E., Harman, D.: TREC: Experiment and Evaluation in Information Retrieval. MIT Press (2005)
21. Bates, M. J.: The Design of Browsing and Berrypicking Techniques for the Online Search Interface. http://www.gseis. ucla.edu/faculty/bates/berrypicking.html (1989)