

# The Dynamics of Interactive Information Retrieval, Part II: An Empirical Study From the Activity Theory Perspective

Yunjie Xu and Chengliang Liu

School of Computing, National University of Singapore, 3 Science Drive 2, Singapore, 117543.

E-mail: xuyj@comp.nus.edu.sg

**Human information-seeking behavior is complicated. Activity theory is a powerful theoretical instrument to untangle the “complications.” Based on activity theory, a comprehensive framework is proposed in Part I (Y. Xu, 2007) of this report to describe interactive information retrieval (IIR) behavior. A set of propositions is also proposed to describe the mechanisms governing users’ cognitive activity and the interaction between users’ cognitive states and manifested retrieval behavior. An empirical study is carried out to verify the propositions. The authors’ experimental simulation of 81 participants in one search session indicates the propositions are largely supported. Their findings indicate IIR behavior is planned. Users adopt a divide-and-conquer strategy in information retrieval. The planning of information retrieval activity is also partially manifested in query revision tactics. Users learn from previously read documents. A user’s interaction with a system ultimately changes the user’s information need and the resulting relevance judgment, but the dynamics of topicality perception and novelty perception occur at different paces.**

## Introduction

Human information-seeking behavior is complicated. The extant literature has approached this issue from the cognitive perspective (Harter, 1992; Ingwersen, 1992, 1996), the behavioral perspective (Bates, 1990; Xie, 2000), and the multisession perspective (Kuhlthau, 1993; Lin & Belkin, 2005; Vakkari, 2001). These perspectives have made great contributions to our understanding of information-seeking behavior. In Part I (Xu, 2007) of this report, we proposed an integrated framework based on activity theory (Leont’ev, 1978; Vygotsky, 1978) for interactive information retrieval (IIR) behavior. The advantage of activity theory lies in its implications on the underlying mechanisms governing the interaction between users’ cognitive state and their manifested behavior in an information-seeking activity. Particularly, it helps explain the transition of a user’s cognitive state

as well as that of a user’s query specification behavior and the reciprocal effect between the two.

Although theoretically appealing, the viability of activity theory in IIR behavior demands empirical testing. The purpose of this article is to report the results of such an empirical study. Because the application of activity theory to IIR is still in its initial stage, our empirical study should be regarded as exploratory in nature. To make this article self-contained, we shall first briefly summarize the propositions of IIR behavior formulated in Part I based on activity theory. After that, we report on an exploratory study including both methodology and data analysis. We conclude with a discussion and the implications of our findings.

## An Activity Theory-Based Framework

Activity theory takes activity as the unit of analysis (Leont’ev, 1978; Vygotsky, 1978). According to that, an interactive information retrieval activity is the totality of the user’s cognitive state, the information retrieval (IR) system and documents, and the interaction between the two. Activity theory also mandates that the dynamics of the above elements should be understood from a developmental and integrative perspective, that is, the state of each element is affected by other elements as well as the history of all elements.

The activity theory-based IIR framework (referred to as “the framework” hereafter) first defines a set of related concepts. It posits that a user’s information problem engenders a general information need which can be decomposed into subneeds. Some subneeds are directly related to the obtaining and use of information, termed *productive subneeds*, whereas others are supportive (we ignore supportive subneeds for simplicity). A general information need is a collection of productive subneeds. When users formulate queries (including query revision) to find documents for subneeds, they use a series of *query revision tactics* (Bates, 1990). The framework also posits that the manifested behavior in an information-seeking process can be decomposed into *sessions*, sessions into *retrieval rounds* (Xie, 2000), and retrieval rounds into *query formulation*, *relevance judgment*, and *document learning* (Vakkari, 2001).

Received March 29, 2006; revised June 14, 2006; accepted July 23, 2006

© 2007 Wiley Periodicals, Inc. • Published online 21 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20574

With these concepts defined, and based on the basic tenets of activity theory, a set of four propositions are proposed with the motivation to describe more precisely the patterns that govern users' cognitive activity (i.e., how do they plan their IIR activity), the effect of cognitive activity on manifested behavior (i.e., query revision), the effect of retrieved documents on manifested behavior, and dynamics of relevance judgment. Collectively, we seek to provide a description of mechanisms that governs the interaction between users' cognitive state and retrieved documents. We shall briefly recap the rationale for the four propositions in the following sections.

Proposition 1 describes the pattern of users' cognitive activity, that is, how users plan their search activity. Consistent with activity theory and prior conceptualizations in IIR (Vakkari, 2001; Xie, 2000), the framework suggests that when a general information need consists of multiple subneeds that cannot be fully specified in one query, users tend to break it down and plan the sequence of searches, although this plan might be disturbed and modified by the actual outcome of user–system interaction. Therefore, users adopt a divide-and-conquer strategy in information seeking:

Proposition 1. Users adopt a divide-and-conquer strategy in IIR. (a) They divide their information need into subneeds, and (b) subneeds are scheduled in the search process.

Because users' subneeds are planned, and because query specification tactics steer queries through a corpus to satisfy subneeds, the use of tactics should also exhibit the divide-and-conquer strategy. That is, users will work on one subneed at a time and read more detailed aspects of it. Reflecting on the use of query revision tactics, such focused search implies the use of deepening tactics in query revision. When sufficient information has been collected, a switch tactic is likely to be employed to move to another subneed. The framework proposes that:

Proposition 2. Users will alternate deepening tactics and switching tactics in IIR.

The use of narrower terms (Bates, 1990; Efthimiadis, 2000; Vakkari, Pennanen, & Serola, 2003) is a good indication of the use of deepening tactics, that is, an indication that a user wants to explore more of the current subtopic, rather than moving on to a new one. Tactics like “the use of related terms” (to replace or add a term that is not a broader or narrower term of a term in the previous query, but relates to it as a cause, a consequence, a concurring event, or a parallel phenomenon, etc.) is more ambiguous in purpose. It might be used for either deepening the current topic or switching to a new topic. In this study, we use the “use of narrower terms” and “switch to a new topic” as two representative tactics for deepening and switching tactics, respectively.

Proposition 2 describes the impact of cognitive planning on the use of tactics. Therefore, it describes the *externalization* process.

Proposition 3 describes the impact of retrieved documents on the use of tactics, that is, the *internalization* process. When documents retrieved are read, users need to judge whether the previous query was on the right track.

They may find that they have used a wrong terminology or a wrong word form. To pull the query back to the right track, some tactics such as making a morphological change of query terms, using the right terminology, and using synonyms are likely to be used to ensure the accuracy of a query. We term these tactics the *topicality-preserving* tactics. Terms for these tactics are more likely from on-topic documents. If the previous query is effective in retrieving documents, users' focuses for the next search are likely to shift to a more detailed aspect of the current subtopic or a new subtopic, a situation we define as *novelty-seeking*. Some tactics such as the use of narrower terms, use of related terms, and switching subtopics are more likely employed if a user wants to learn new information. The terms used to revise the query are more likely from the novelty aspect of documents read. Therefore:

Proposition 3. (a) Novelty-seeking tactics such as the use of narrower terms, use of related terms, and switching topics are more likely to draw terms from the novelty aspects than from the topicality aspect of documents read. (b) Topicality-preserving tactics such as the morphological change of query terms, use of synonyms, and replacing wrong terminology are more likely to draw terms from the topicality aspects than from the novelty aspects of documents read.

Based on this framework, the internalization and externalization processes not only reciprocally affect cognitive state and query specification, but also change the relationship between them. Particularly, as a user's general information need is satisfied over time, the marginal contribution of documents encountered later decreases. This decrease is more due to the reduced novelty than topicality:

Proposition 4. In an IIR session, the decrease of perceived document novelty is faster than that of document topicality.

## Methodology

We conducted an exploratory study to test the propositions. Note the propositions are at a conceptual level and might be tested in various ways. What we did in this study was to test them in one way. Therefore, this study was intended to be an exploratory study rather than a confirmatory one. We sought to explore and describe some patterns of IIR behavior that were consistent (or contradictory) to the propositions.

Activity theory does not prescribe any particular research methodology although field studies are used most of time. Engeström (1987) proposed a few guidelines for empirical study. Engeström highlighted that before observation, it is important to delineate an activity by the motive that drives it. To observe an activity, he stressed attention to object history, participant history, and interaction history, which links the former two. Borlund (2003) proposed the use of experimental simulation for IIR study. She suggests that experimental simulation has the advantage to maintain a cognitively individual information-need interpretation as in real life and an experimental control for a researcher. Because it is possible to define an IIR activity with a clear boundary, such as an IIR session with a well-defined information need, and because

such activity does not involve extensive social interaction, we believe experimental simulation is a suitable method.

Basically, in our experimental simulation, we invited participants to participate in an IIR activity. We recorded their question revision history, document evaluation history, and query specification history. We also asked users to indicate their tactics in the IIR process. Such observations gave us a window to understand the structure and dynamics of the users' information need and subneeds, query specification, relevance judgment, and their interactions.

### Search Task

Eighty-one (81) first-year undergraduate students participated in the IIR experimental simulation. They were paid \$15 for their participation. The information-seeking task was described with the following search statement:

Assume you are taking a health education course. You will be asked to search online for documents about "the relationship between mobile phone radiation and health." Documents addressing the following questions considered relevant are:

- *Does use of mobile phones pose radiation threats to users' health?*
- *Why are there such or no such radiation threats to health?*
- *What is the proper way to use a mobile phone to protect your health from radiation?*

You will take an **online quiz** (10 questions) based on the knowledge you learned in the search process. Those who score well ( $\geq 80$ ) will be entered in a lottery to win \$50. If nobody scores well, the person with the highest score will get the prize.

In experimental simulation for IIR research, one important requirement is to ensure users' involvement (Borlund, 2003). We consider the choice of topic as suitable because 91% of the population in the experiment locus use mobile phones, and most participants expressed an interest in the topic of informal chats. The complexity of the topic was also adequate because it involved multiple related subneeds that ask "what," "why," and "how" questions. These subneeds were unlikely to be addressed in a single query. The complexity of the search statement allowed us to explore the potential division and planning of subneeds. The use of the online quiz and monetary incentive was to (a) further motivate users to take the task seriously, (b) create a task that

requires learning so that we could observe the impact of documents on the users' cognitive states, and (c) give a clear motive and end to the activity.

### The Corpus, Search Engine, and Activity

We developed our own search engine based on open-source packages for vector space models. The search engine ran on a Web server and the interface was Web-based.

The corpus for this experiment was 620 online documents collected from Google.com. To ensure the corpus had both relevant and irrelevant documents, more than 10 queries of various accuracy were used such as "mobile phone radiation health protection," "mobile phone," and "health-mobile-phone" ("-" means "exclude" at Google.com). For each query, links from the first 30 pages of Google.com returns were all downloaded, but duplicates, navigation pages with all links and no content, advertisements, and other non-article Web pages were excluded. In total, we had 270 documents based on more "accurate" queries, and 350 based on "ill-specified" queries. Documents were then cleaned, indexed, and supplied to our search engine.

The search engine did not allow operators (e.g., and, or, not, -) in the specification of the question. The interface of the search engine was very simple. It started with a simple search box like that of Google.com, and returned a list of 10 document titles linked to the actual documents (Figure 1). Users needed to read and evaluate all the returned documents in order. Each document was evaluated based on its topicality, novelty, and overall usefulness. Usefulness and relevance criteria were measured on an 8-point Likert-scale. There were four anchors on the scale (e.g., 0 = *Useless*, 1 = *A bit useful*, 4 = *Satisfactory*, and 7 = *Essential*). Although the conventional use of the Likert-scale uses a 7-point rating system, Tang, Shaw, and Vevea (1999) found that the optimal scale for relevance measure is about 7.7. We added the zero-anchor to more clearly indicate the uselessness (or the like) of a document. These relevance criteria were defined in the experiment instructions as follows:

#### How to Evaluate Documents?

For each query, the search engine returns 10 documents. Please read the documents one by one, and evaluate each one in terms of whether it is **On-topic**, **Novel (provides new knowledge)**, or **Overall Useful**.

**Overall Useful:** A document is overall useful if it makes a major contribution to your information need and you expect it to substantially **contribute to your quiz grade and**

Title	NOVEL	ON-TOPIC	USEFUL
<u>Mobile Phone</u>	NOVEL [dropdown]	ON-TOPIC [dropdown]	USEFUL [dropdown]
<u>Don't give mobile phones to the under-9s</u>	NOVEL [dropdown]	ON-TOPIC [dropdown]	0-USELESS 1-A BIT USEFUL 2 3 4-SATISFACTORY 5 6 7-ESSENTIAL
<u>It's a mobile world</u>	NOVEL [dropdown]	ON-TOPIC [dropdown]	
<u>Using GSM - Health &amp; Environment Issues - FAO's</u>	NOVEL [dropdown]	ON-TOPIC [dropdown]	

FIG. 1. Our search box.

**you try to memorize its content.** You may assign it a score ranging from 0 (No Contribution), 1 (Very Low Contribution) to 7 (Very High Contribution).

**On-topic:** A document is on-topic if its main content is **related** to the subject area of your query, i.e., you would classify this document into a category of documents labeled with your query. However, an on-topic document is not necessarily useful. You may assign it a score from 0 (Totally Off-topic), 1 (Marginally On-topic), to 7 (Substantially On-topic).

**Novel:** A document is novel if it provides **new knowledge** to you, i.e., you didn't know the information before reading this document. Again, a novel document is not necessarily useful. You may assign it a score ranging from 0 (Nothing New), 1 (Very Little New Knowledge) to 7 (Very Much New Knowledge) based on how much it is novel.

We used the term "usefulness" to represent the situational relevance of a document. Usefulness is not a relevance criterion, but an overall measure of situational relevance (Fitzgerald & Galloway, 2001). Topicality and novelty are relevance criteria (Xu & Chen, 2006). After evaluating all of the returned 10 documents, users needed to specify a new query or revise the current one in a search box at the bottom of the page.

After users had evaluated the returned 10 documents and revised a query, they were led to another page to explain their use of query specification tactics. For every query revision, they were asked to pick all tactics they applied to the revision from a pool of seven tactics. These seven tactics included use of broader concept/terms (Broader), use of narrower concept/terms (Narrower), use of related concept/terms (Related), use of synonyms (Synonym), correct wrong terminology (WT), morphological changes (MC), and switching of subtopics (Switch) (The definitions of these tactics are given in Part I of this article. They were also included in the instructions to the participants). All terms added must be explained. Term deletion was also recorded, but not of interest to this study. In addition, for the added terms, users needed to indicate their source: from the documents read or from other sources, such as the users' own background knowledge. Correction of spelling errors was not included as a query specification tactic because we considered it as purely supportive. Such a problem can largely be solved technically. In our system, we allowed users to cancel a round if a spelling error was discovered. After explaining the query revision, a second set of 10 documents was retrieved from the remaining corpus for evaluation. Documents retrieved in each round were not repeated.

### *Experimental Process*

The experiment was carried out in a computer laboratory in four sessions with about 20 users each. We first introduced the search topic, and asked users to fill out a preexperiment questionnaire with their basic demographics and three subjective questions on their perceived knowledge of the topic. Then we demonstrated the search engine with an irrelevant topic (car accidents) and explained the document evaluation criteria and the query specification tactics. Users were asked to try the system to gain hands-on experience with an

irrelevant topic using a different corpus. This training session took about 30 minutes. Users were then directed to the main search interface. They were asked to evaluate at least four rounds of documents before taking the quiz. After taking the quiz, we asked them to fill out a postexperiment questionnaire to evaluate their perceived knowledge again. After the experiment had been completed, the prize was awarded to the winner. Users took about 1.5 to 3 hours to finish the whole process.

### *Data Analysis*

Among the 81 users participating in the experimental simulation, 31 (38.2%) were men and 50 (62.8%) women; the average age was 19.66 years ( $SD = 1.42$ ). They were experienced users of Internet search engines ( $M = 5.01$  years,  $SD = 1.97$ ) and mobile phones ( $M = 3.06$  years,  $SD = 1.7$ ). Before the experiment, their self-evaluated knowledge level was 3.35 ( $SD = 1.28$ ) based on a 7-point Likert scale (1 = *have very little knowledge*, 7 = *is very knowledgeable*). After the experiment, the average rose to 5.5 ( $SD = 1.28$ ), indicating that learning did occur during the information-seeking process. Most users searched for four rounds; only 9 (11%) searched for five rounds. To make the rounds comparable, we report only the analysis for the first four rounds, which consisted of 324 ( $81 \times 4$ ) queries. We left the fifth round out because given the small number of observations the statistical estimates (e.g., means, variance) may not be stable. Moreover, a valid statistical comparison between rounds requires a relatively large sample size in each group to ensure sufficient statistical power (i.e., high probability to discover a truly existing relationship).

## **Results**

### *Proposition 1: Divide-and-Conquer*

Do users break down their information need into subneeds? Do they schedule them? If a divide-and-conquer strategy is adopted by a user, the query history should reflect a user's scheduling of information subneeds, and we could infer the breakdown of the general information need and the planning of subneeds by observing the use of query terms over rounds.

Table 1 reports the distribution of the top 15 most popular terms over the rounds and their grand totals. In summarizing query terms, we deleted stop words (e.g., a, an, the) and combined morphologically related terms (e.g., phone and phones, health and healthy). There were 109 different terms and the total use of all terms was 1254 times. The top 15 terms accounted for 1075 times (85.7%) of the total use. A consistent pattern in use of terms was that the top four terms (phone, mobile, radiation, health) were the most popular in all rounds and accounted for 71% of all term uses. These terms could be regarded as the key terms to indicate the general topicality as demanded by the search statement. Intuitively, their presence in a document indicates the relevance of the document to the general topic area of the information need.

TABLE 1. Term distribution over the rounds.

Rank	Grand total		Round 1		Round 2		Round 3		Round 4	
	Term	%	Term	%	Term	%	Term	%	Term	%
1	Phone	21.22	Radiation	25.68	Phone	21.00	Phone	21.41	Phone	19.30
2	Mobile	21.07	Mobile	25.00	Mobile	20.70	Mobile	20.45	Mobile	18.40
3	Radiation	18.47	Phone	23.65	Radiation	17.90	Radiation	15.34	Radiation	16.60
4	Health	10.46	Health	11.82	Health	10.30	Health	9.58	Health	10.40
5	Threat	3.66	Threat	6.08	Threat	5.02	Risk	3.51	Protect	5.21
6	Protect	2.75	Pose	1.35	Protect	2.51	Protect	3.19	Threat	1.84
7	Risk	1.45	Harm	1.01	Cancer	1.57	Threat	2.24	Proper	1.84
8	Brain	0.99	Hazard	0.68	Brain	1.25	Brain	1.28	Risk	1.84
9	Proper	0.92	Hand-phone	0.68	Proper	1.25	Cause	1.28	Brain	1.53
10	Effect	0.84	Effect	0.34	Hand-phone	1.25	Child	1.28	Human	1.23
11	Hand-phone	0.84	Mobile-phone	0.34	Cause	0.94	Effect	1.28	Safe	1.23
12	Safe	0.76	Hand	0.34	Against	0.63	Cell	1.28	Cell	1.23
13	Cell	0.76	Relate	0.34	Hand	0.63	Hand-phone	0.96	Effect	1.23
14	Cause	0.76	Cause	0.34	Relate	0.63	Proper	0.64	Precaution	0.92
15	Harm	0.76	Nature	0.34	Hazard	0.63	Safe	0.64	Affect	0.92
Total count	1254		296		319		313		326	

Other terms were used less frequently. However, careful examination of other terms indicated that terms indicative of the health threats of mobile phones (e.g., threat, harm, and cancer) were more frequently used in the earlier rounds than in the later rounds. In contrast, terms that were indicative of protection methods (e.g., protect, proper, safe, and precaution) became more popular in the later rounds. This pattern motivated us to place terms into different groups: general topical terms, threat-related subtopic terms, and protection-related subtopic terms. A set of the most popular six terms was identified from the pool of all query terms for each group (Table 2). Note that these sets of terms were not intended to be comprehensive, but rather representative. We only included terms that were not ambiguous in their group membership. We treated “radiation” as a general topical term rather than a threat-related term because the search topic is about whether radiation poses a threat to the user. A document that does not cover radiation is most likely off the general topic of interest.

Proposition 1a hypothesizes that users divide information needs into subneeds. If threat and protection are two subneeds, then we should observe threat- or protection-related terms to be used in consecutive rounds rather than be distributed evenly over all four rounds. In contrast, general topical terms should be used in all rounds to ensure topicality of a document to the general topic area of the information need. Therefore, subtopic terms are more concentrated than general

TABLE 2. Terms indicative of general topic area, threat, and protection.

Group	Terms
General	Cell, hand phone, health, mobile, phone, radiation
Threat	Cancer, danger, harm, hazard, risk, threat
Protection	Correct, precaution, prevent, proper, protect, safe
Others	All other terms

topic terms. To test that statistically, we defined a concentration measure for terms of a subtopic or the general topic.

For each user, because there are four rounds of query, we denote the rounds as  $\{k = 1, 2, 3, 4\}$ . Assume for a specific topic (subtopic or general topic), there are  $v$  terms in each round; we have  $\{v_1, v_2, v_3, v_4\}$ . Then the “weight center”  $C$  of query terms for this topic can be defined as

$$C = \frac{\sum_{k=1}^4 kv_k}{\sum_{k=1}^4 v_k}$$

which represents the weighted average of the round number for the topic. For example, if some term for a subtopic appears in this way:  $\{0, 1, 2, 0\}$ , the center for this subtopic is  $2.77 (1 \times 0/3 + 2 \times 1/3 + 3 \times 2/3 + 4 \times 0/3)$ .

Take this weighted average of the round number as the center of this topic, the distance from a particular round that also has terms for this topic to this center can be calculated. In the above example, the distance from round 2 to the center is  $0.77 = 2.77 - 2$ . If a topic is concentrated, then such distances will be small. We can define a concentration measure akin to the concept of “variance” in statistics, which is the mean distance square of all rounds to the center. We define our concentration measure  $CM$  as:  $CM = \frac{\sum_{k=1}^4 (k - C)^2 v_k}{\sum_{k=1}^4 v_k}$ . The concentration measure indicates the weighted average distance square of queries of a topic from its center.

Given this definition, we calculated the concentration measure for the general topic, threat, and protection for all users. However, some users did not have classified terms for some subtopics, leading to an unequal number of observations in these topic areas. As reported in Table 3, the center of the general topic was 2.40, which was close to the 2.50, the unweighted center of four rounds. The center of threat was 2.34, and the center of protection was much later at 3.28. General topic terms showed larger mean distance square (1.21) than both threat- and protection-related subtopics (0.38 and 0.11, respectively). Taking the general topic as a

TABLE 3. The center, mean distance square, and significance of difference for three groups.

	Sample size	Center (SD)	M Dist <sup>2</sup> (SD)	p Value for dist <sup>2</sup>
General	81	2.40 (0.25)	1.21 (0.24)	—
Threat	57	2.34 (0.24)	0.38 (0.62)	0.00
Protection	44	3.28 (0.77)	0.11 (0.23)	0.00

benchmark, a *t* test assuming unequal variance indicated that the mean distance squares of threat and protection were much smaller than that of the general topic ( $p < .001$ ). Therefore, subtopics were more concentrated than the general topic. The finding is in support of Proposition 1a.

Proposition 1b hypothesizes that subtopics are scheduled. A *t* test assuming unequal variance indicated that the center of threat was much earlier than the center of protection ( $p < 0.001$ ), suggesting that users focused on threat first, then moved on to protection.

To further understand users' divide-and-conquer strategy, we plotted the total number of terms used for each group (Figure 2). As in Table 1, overall, general terms were more popular than subtopic terms in query formulation. What was more interesting was the trend. Table 4 reports the means of number of terms for all rounds and paired *t*-test results ( $N = 81$ ) among the means. For general topic terms, the first round was significantly higher than the later rounds, but the differences between Rounds 2–4 were insignificant. This observation indicates that users conducted a "topic overview" before delving into subtopics. For threat-related terms, there was no difference among all rounds, suggesting that the interest in threat was roughly constant over all rounds at the group level. For protection-related terms, there was a clear upward trend that most users selected this subtopic in the later rounds of retrieval. Except that Rounds 2 and 3 were not

TABLE 4. The mean number of terms for three groups over rounds and the significance of their difference.

	Round	M	SD	1	2	3
General	1	3.17	0.77			
	2	2.81	0.90	0.00		
	3	2.67	1.00	0.00	0.23	
	4	2.65	1.12	0.00	0.24	0.98
Threat	1	0.28	0.45			
	2	0.33	0.47	0.50		
	3	0.30	0.46	0.86	0.60	
	4	0.23	0.53	0.50	0.18	0.38
Protection	1	0.01	0.11			
	2	0.19	0.45	0.00		
	3	0.22	0.42	0.00	0.55	
	4	0.38	0.58	0.00	0.03	0.06

significantly different, other differences were all significant (the difference between Rounds 3 and 4 was very close to significance with  $p = .056$ ). Overall, the group level pattern indicates that users scheduled their subneeds. In addition, it indicates the special role of the first query as an indicator of general topic interest.

A clearer picture can be obtained if we look at the terms added to the previous queries over the rounds, because added terms indicate the user's shift of interest. From Round 2 to Round 4, there were 404 added terms, comprised of 94 different terms. We compared the percentages of added terms for the three groups (Figure 2b). General terms were still the most popular over all rounds to maintain the general topicality of query. Meanwhile, users added less threat-related terms in later rounds, but more protection-related terms. At Round 2, the percentage of threat terms was significantly higher than that of protection ( $\chi^2$  test,  $p = .006$ ). At Round 4, the difference was the other way around ( $\chi^2$  test,  $p = .02$ ). This pattern again indicates that users scheduled threat before protection.

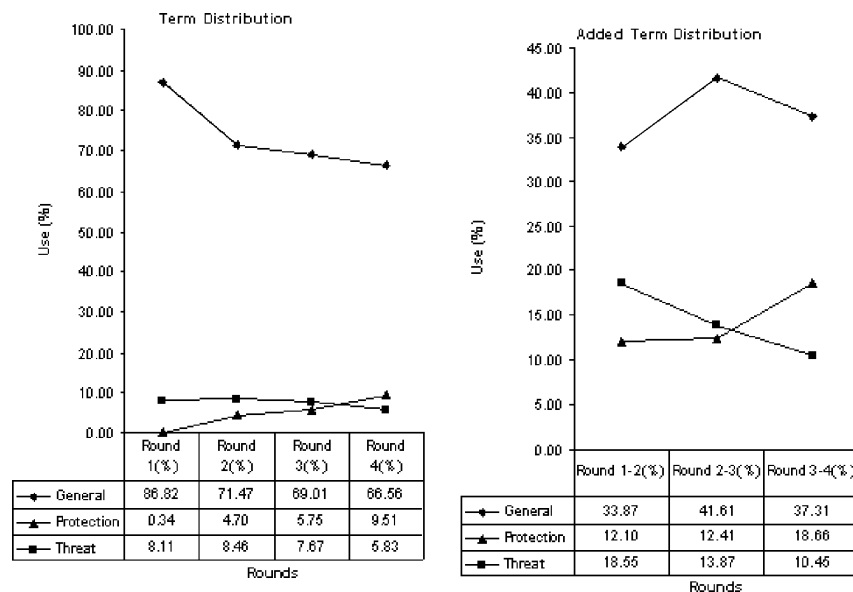


FIG. 2. Plot of the total number of terms used and number of added terms for each group.

However, the particular ordering here should be interpreted with caution. A plausible reason that threat appeared before protection, but not the other way around, is that the nature of the search task required users to understand the threats of mobile phone use before taking precautionary measures. Had the subneeds of the search task not required temporal ordering (imagine if the search topic was to study a certain period of French and German history), users might start with any subtopic and not demonstrate a systematic pattern at the group level. Moreover, the description of the search topic put the understanding of threat before protection means, which could also influence participants' planning of the search activity. Therefore, we conclude that users did schedule their subneeds, but the actual ordering of subneeds was subject to the influence of the topic nature and the search instruction.

Overall, we consider the evidence supportive of the Proposition 1a. The support to Proposition 1b is only conditional. Future research should identify the factors and conditions that affect the scheduling of subneeds in a search plan.

*Proposition 2: Use of Query Revision Tactics*

Proposition 2 hypothesizes that users alternate deepening tactics and switching tactics in IIR. Over the four rounds, there were 322 uses of tactics, with 1.33 tactics per revision. Because there was no pre-established classification of which tactics are for deepening and which are for switching, we consider Narrower (use of narrower terms) as a deepening tactic, and Switch a switching tactic. Other tactics are less clear-cut. For example, a deeper investigation indicates that users tended to confuse Switch and Related: when they added or substituted in, say, protection-related terms for threat-related terms, some indicated Related, some Switch. This confusion arose because the two tactics are not exclusive of each other. It also raises the question of whether the existing classification scheme for tactics is adequate for our purpose. It seemed that a simple classification with “deepening on the current subtopic” and “switching to a new subtopic” as the two major tactics might be more useful in future studies. Therefore, in the following analysis, we consider Narrower as a representative of deepening tactic, but not Related. Table 5 reports the number of uses for all tactics in percentages.

TABLE 5. Percentages of query specification tactics.

Tactics	Revision 1 (%)	Revision 2 (%)	Revision 3 (%)	All rounds (%)
Narrower	35.29	18.18	27.27	25.50
Related	19.61	23.64	16.36	20.79
Switch	14.71	25.45	24.55	25.25
Morphological change	3.92	5.45	6.36	3.71
Synonyms	14.71	17.27	8.18	11.88
Wrong terminology	3.92	1.82	3.64	3.47
Broader	7.84	8.18	13.64	9.41
Total frequency	102	110	110	322

Similar to prior studies (Efthimiadis, 2000; Vakkari et al., 2003), we also found Narrower, Related, and Switch to be the most often used tactics (all above 20%). Moreover, our analysis reveals that in the first revision, the use of narrower terms was the dominant tactic ( $\chi^2$  test indicates the percentage of Narrower is significantly higher than all other tactics with  $p < .001$ ). This is consistent with what we observed in term distribution, where the first round had more general topic terms. At the group level, it seems that users focused on the general topic in the first round, and delved into a subtopic in the second round. For the second revision, Switch and Related were most popular, and Switch had a significantly higher percentage than Narrower ( $\chi^2$  test,  $p = .048$ ). For the third revision, Narrow and Switch were the most popular, but the difference was insignificant ( $\chi^2$  test,  $p = .50$ ). In the first two revisions, the result was consistent with Proposition 2 at the group level. The pattern did not continue to the third revision, possibly because of the different speeds at which the information need was satisfied for different users. At group level, Table 5 also indicates that the most probable four sequences of tactics over rounds at the group level were: Narrower—Switch—Narrower ( $p = .0245$ ), Narrower—Related—Narrower ( $p = .0228$ ), Narrower—Switch—Switch ( $p = .0220$ ), and Narrower—Related—Switch ( $p = .0205$ ). The close probabilities of the four sequences indicate that although the Narrower—Switch—Narrower sequence was most popular, users' tactic sequences were not limited to it.

Notice the above analysis was done at the group level rather than the individual level. An individual level analysis would require us to count all possible tactic sequences over the four rounds. However, the most popular three-tactic sequence occurred only 5 times (Narrow—Switch—Switch), which was not statistically informative. We also looked at the frequency of two-tactic sequences. However, because there might be multiple tactics in one revision, we ignored all tactics which involved only stop words (e.g., replace the term “for”) and ended up with 162 two-step patterns. If there were different tactics (e.g., a query has both Narrower and Broader tactics) in one query revision, such cases were also left out (29 cases) (The same “cleaning” processing was used for the counting of three-tactic sequences mentioned above.). For the remaining 133 patterns, 106 started with Narrower, Related, or Switch. Table 6 reports the distribution of the major tactic sequences over all rounds.

Sequences starting with Narrower, Switch, and Related were most often used and Switch was used significantly more often than Narrower ( $\chi^2$  test,  $p = .02$ ). After Switch, the use of another Switch was significantly more likely than Narrower ( $\chi^2$  test,  $p = .013$ ). Therefore, if we take Narrower as the starting tactic (Narrower was the most popular tactic for the first revision), this table indicates that at the two-tactic sequence level, if we sequentially connect two two-tactic sequences back to back, Narrower—Switch—Switch was the most frequent sequence, followed by Narrower—Switch—Related and Narrower—Switch—Narrower. This result is different from the individual tactic analysis in Table 5

TABLE 6. Conditional probability of one tactic after another.

Revision (i-1)	Revision (i)	Frequency	Conditional probability (%)
Narrower	Narrower	8	18.6
	Related	9	20.9
	Switch	14	32.6
	Others	12	27.9
Related	Narrower	6	25.0
	Related	7	29.2
	Switch	6	25.0
	Others	5	20.8
Switch	Narrower	11	28.2
	Related	6	15.4
	Switch	18	46.2
	Others	4	10.3

in that Switch is now more likely to be followed by another Switch than by Narrower in the third query revision. This difference might be due to the elimination of some observations. Another plausible reason is that users were more easily satisfied with documents retrieved in later rounds so that the need for further narrowing down was reduced; instead, they hastily moved on to another subtopic.

In summary, the evidence indicates the alternate use of deepening and switching tactics in the first two query revisions. However, the sequence in later revisions was inconsistent between group level analysis and two-tactic sequence analysis. Notice the later revisions are affected by all previous revisions and their outcomes; hence, it is more difficult to find a clear pattern in the later rounds. We consider the support for Proposition 2 to be weak. Further research is needed with a better tactic scheme and a longer observation period.

*Proposition 3: Terms for Tactics*

This proposition hypothesizes that users learn topicality and novelty terms from documents read. Topicality terms are more likely to be learned from the topicality aspect of documents and novelty terms from the novelty aspect. However, how do we know if a term is regarded as a topicality term or a novelty term by a user? Because we did not explicitly ask users to indicate it, we inferred term use from the

tactics involving it. Tactics such as Narrower, Related, and Switch are more likely used to obtain new information. Therefore, their terms are more likely used for novelty seeking. In comparison, morphological change (MC), use of synonym (Synonym), and replacement of wrong terminology (WT) are tactics to bring back or maintain the current topicality. Terms for such tactics are regarded for topicality preserving. However, the membership for Broader is unclear. We left it as a separate group.

How do we calculate the probability of a term from a document? Although a few techniques are available, we used the language model (Ponté & Croft, 1998) because of its focus on query production, that is, its explicit modeling of the probability of a query term being “drawn” from a document. Because there were 10 documents in each round, we merged them into one: For each document, we ignored the semantic relationship among terms and treated them as a bag of words, then added up the frequency for each term from all 10 documents weighted by the topicality or novelty score of a document. In this way, a merged document was created. When the weight of documents used in the merging process was a topicality score, we called the merged document the topicality profile; otherwise, it was a novelty profile. In the topicality profile, highly on-topic documents had a higher weight and in the novelty profile, highly novel documents had a higher weight. Notice the two profiles contained the same bag of words. The only difference was how the terms were weighted. With these two profiles ready, we could calculate the probability of each query term  $w$  being “drawn” from the topicality profile  $P^T(w)$  and its probability from the novelty profile  $P^N(w)$ . The detailed formula for calculation is reported in the Appendix.

Including Round 5, we had, in total, 434 added terms (excluding stop-words). Of these terms, 312 were from the documents read. We excluded those terms from users’ background knowledge. Table 7 reports the simple average of probabilities of terms for various tactics.

For novelty-seeking tactics, added terms were more likely from the novelty profile ( $p = .0054$ ) than from the topicality profile ( $p = .0046$ ). This pattern was consistent for all three tactics. Paired  $t$  test shows that the differences were all significant ( $p < .05$ ). The implication is that terms for these three tactics were based more on the novelty

TABLE 7. The novelty and topicality probability of terms.

Goal	$N$	$P^N(w)$	$SD$	$P^T(w)$	$SD$	$T$ -Value	$p$ -Value
Novelty seeking	<b>230</b>	<b>0.0054</b>	<b>0.0091</b>	<b>0.0046</b>	<b>0.0084</b>	<b>3.75</b>	<b>0.00</b>
Narrower	76	0.0051	0.0089	0.0047	0.0083	2.31	0.02
Related	80	0.0067	0.0108	0.0056	0.0101	2.20	0.03
Switch	74	0.0044	0.0071	0.0035	0.0060	2.49	0.01
Topicality seeking	<b>67</b>	<b>0.0060</b>	<b>0.0091</b>	<b>0.0054</b>	<b>0.0083</b>	<b>1.45</b>	<b>0.15</b>
MC	8	0.0044	0.0052	0.0044	0.0051	0.05	0.96
Synonym	47	0.0070	0.0099	0.0062	0.0091	1.31	0.20
WT	12	0.0032	0.0069	0.0029	0.0064	1.83	0.09
Others	15	0.0043	0.0046	0.0044	0.0051	-0.20	0.85
BROAD	15	0.0043	0.0046	0.0044	0.0051	-0.20	0.85

Note. MC = morphological changes; WT = wrong terminology.

perceptions of the documents read, than on the topicality perceptions. Therefore, Proposition 3a is supported. The picture is less clear for topicality-preserving tactics. Except Synonyms, MC, and WT were rarely used, suggesting that they act more like exception handling supportive tactics in query specification. The probability differences for all three tactics were insignificant, which does not support Proposition 3b. Similarly, terms for Broader have no clear topicality or novelty orientation as anticipated. Overall, we conclude that users drew terms from the novelty aspect of retrieved documents for novelty-seeking tactics, but for topicality-preserving tactics, there was no significant difference.

*Proposition 4: Dynamics of Relevance Dimensions*

Finally, as users' information need and subneeds are satisfied over time, we hypothesized that document novelty decreases faster than document topicality. To verify that, for each user, we averaged the topicality, novelty, and relevance scores (0–7 is normalized to 0–1) for each round of documents evaluated. Then, we averaged them for all users. Such measures can be regarded as accumulated precision scores for each round (Kekäläinen & Järvelin, 2002). Figure 3 illustrates the change of topicality, novelty, and overall relevance over four rounds.

Users' novelty perception of documents decreased sharply from Round 1 (0.44) to Round 2 (0.35). However, the decrease in topicality was milder (−0.05). From the second round on, topicality remained steady, but novelty continued to decrease and then leveled off. Although the magnitude of the difference in the trends of topicality and novelty is small, this pattern is consistent with Proposition 4 at the aggregated level.

We applied trend analysis (Kirk, 1995) for the novelty and topicality perception of all users. Trend analysis allows us to detect linear, quadratic, or cubic trends over time. Novelty perception had significant linear ( $t = -4.76, p < .005$ ) and quadratic ( $t = 2.57, p = .01$ ) trends, indicating that it decreased first, but the decrease was gentler or even reversed later. The cubic trend was insignificant. Topicality perception did not show any significant trend. The overall relevance pattern was more parallel to novelty than to topicality.

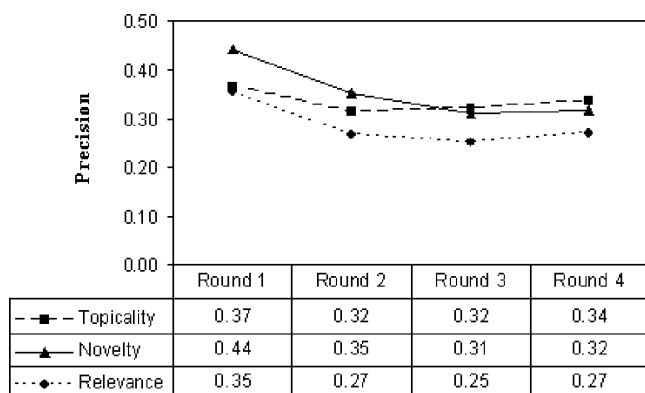


FIG. 3. The change of topicality, novelty, and overall relevance over four rounds.

Overall, the difference in significance of the downward trend between topicality and novelty supports Proposition 4.

**Discussion and Implications**

The purpose of our experimental simulation was to verify the activity theory-based IIR framework and the propositions based on it. The findings of the empirical study are generally consistent with the theory except for Proposition 3b. In addition, the support for Proposition 1b should be regarded as conditional. We summarize the findings in Table 8.

*Limitations*

Before we discuss the implications of this study, some limitations should be noted. First, this study adopts an experimental simulation approach. We forced participants to evaluate all documents, which was obviously not realistic. More naturalistic user studies may help verify our findings. Second, in the experimental context, we used a single task with a private IR system and corpus. More realistic information needs and interaction with more complicated IR systems might offer additional insight. Third, we focused only on a single-session context. Investigating multisession scenarios can enrich and expand this framework. In this single-session context, we also ignored supportive information retrieval behaviors, which could be interesting in other contexts. Fourth, our propositions can be tested in other ways. The data-checking points employed could limit our understanding of them. Fifth, the limited support for Proposition 1b and Proposition 2 calls for better research design in the future. Finally, this study involves many measurements of users' subjective perceptions. Measurement validity and measure quality need to be verified in future research.

*Theoretical Implications*

This study represents only a first attempt to apply activity theory to IIR and interactive information search (IIS) research. There are other aspects of activity theory (e.g., the sociohistorical perspective) that are relevant to IIR and IIS research, and the main concepts of activity theory (e.g., concepts related to users' cognition) can be measured and applied to IIR and IIS research in other ways. For example, we have only looked at the planning activity of users' cognition. Given the limited exploration we have done in this study, the theoretical implication of this study is multifold. First, it suggests that activity theory is a viable theoretical lens to view IIR behavior. It provides an alternative perspective to the current cognitive and behavioral perspectives and advocates the totality of activity as the unit of analysis. The advantage of the activity theory based framework is to uncover the reciprocal effects between users' cognitive states and manifested behavior. The focus on interactions through internalization and externalization can be regarded as the hallmark of this framework. To user-centered researchers, this perspective calls for more in-depth behavioral

TABLE 8. Summarization of the exploratory study.

Proposition	Data checking points	Findings
<b>Proposition 1.</b> Users adopt a divide-and-conquer strategy in IIR. (a) They divide their information need into subneeds. (b) Subneeds are scheduled in the search process.	<ul style="list-style-type: none"> <li>• The distribution of all query terms over the rounds</li> <li>• The distribution of groups of added query terms corresponding to general topic area and subneeds over the rounds</li> <li>• The concentration measure of all groups</li> </ul>	<ul style="list-style-type: none"> <li>• General terms are dominating over the rounds</li> <li>• Users start with “topic overview”</li> <li>• Subtopic terms are concentrated in consecutive rounds</li> <li>• Addition of subtopic terms exhibit trends over rounds, but this pattern might be a result of search topic instruction</li> </ul>
<b>Proposition 2.</b> Users alternate deepening tactics and switching tactics in IIR.	<ul style="list-style-type: none"> <li>• The distribution of query specification tactics over the rounds</li> <li>• The conditional distribution of tactics in two consecutive rounds</li> </ul>	<ul style="list-style-type: none"> <li>• Over the rounds, the use of all tactics follows a few major sequences</li> <li>• The use of two-tactic sequences also displays some major sequences</li> <li>• Alternate use of deepening and switching tactics is one of the main sequences, but not the only one</li> <li>• Alternate use of deepening and switching tactics is clearer in earlier rounds, but not in the later rounds</li> </ul>
<b>Proposition 3a.</b> Novelty-seeking tactics such as use of narrower terms, use of related terms, and switching topic are more likely to draw terms from the novelty aspects than from the topicality aspect of documents read.	<ul style="list-style-type: none"> <li>• The probability of terms used for different tactics being drawn from the novelty profile based on the language model of the previous round</li> </ul>	<ul style="list-style-type: none"> <li>• Terms for novelty-seeking tactics (Narrow, Related, Switch) are significantly more likely to be drawn from a novelty profile than from a topicality profile</li> </ul>
<b>Proposition 3b.</b> Topicality-preserving tactics such as morphological change of query terms, use of synonyms, and wrong terminology replacement are more likely to draw terms from the topicality aspects than from the novelty aspects of documents read.	<ul style="list-style-type: none"> <li>• The probability of terms used for different tactics being drawn from the topicality profile based on the language model of the previous round</li> </ul>	<ul style="list-style-type: none"> <li>• The difference of probabilities of terms for topicality-preserving tactics (MC, Synonym, WT) being drawn from novelty and topicality profiles are insignificant</li> </ul>
<b>Proposition 4.</b> In an IIR session, the decrease of perceived document novelty is faster than that of document topicality.	<ul style="list-style-type: none"> <li>• The average topicality and novelty over rounds</li> </ul>	<ul style="list-style-type: none"> <li>• The average topicality and novelty perceptions decrease, but novelty perception decreases at a faster pace</li> </ul>

Note. IIR = interactive information retrieval.

studies that probe the micro-level transitions of users’ cognitive states and their manifested interaction with systems.

Second, with activity as unit of analysis, this study suggests that users’ IIR behavior exhibits a divide-and-conquer strategy even within a session. The actual breakdown of users’ information needs, however, might hinge on the nature of the information need and problem situation. Therefore, if a query revision pattern is what is to be discovered or predicted, it cannot be separated from users’ information need and the problem situation. Components of an information need might be temporally ordered. Temporal order provides a natural cue in predicting query formulation over time. Kuhlthau’s (1993) six stages of information seeking can be regarded as a general description of temporal order in scheduling subneeds. However, temporal breakdown is not the only way: Information need could be broken down into causally related sequences. It could also be ordered by the importance of subneeds to users, or even affected by the search problem description. Understanding of these factors seems to be critical to the prediction of users’ scheduling of subneeds.

Third, users’ manifested interaction with a system, that is, query specification, is also patterned. First, a query must maintain its relevance to the general topicality of the information need. However, just hitting the general topic area is not enough. Terms representing subneeds are intentionally added to users’ queries. Users normally deepen on a subtopic area first. Tactics like Narrower are employed. If the subneed is satisfied, users switch to another subneed. Such alternative use of deepening and switching tactics reflects a divide-and-conquer strategy in resolving information need.

Fourth, different tactics make use of different terms that users learned from the previous round. Some tactics seem to be more novelty seeking in nature, such as Narrower, Related, and Switch, whereas others (MC, WT, Synonym) are used to maintain the current topicality and draw the query back on the right track. Novelty-seeking tactics draw terms from the novelty aspect of previous documents whereas topicality-preserving tactics exhibit no particular preference.

Fifth, consistent with earlier findings by Vakkari (2001), our findings suggest that the dynamics of different dimensions

of users' relevance judgment is in fact different. Topicality judgment is more stable than the novelty judgment. It is easier for an IR system to satisfy users' topicality requirements, but more difficult for novelty. Therefore, as suggested by Xu and Chen (2006), quantifying novelty is an important direction for future research.

Finally, rather than offering a comprehensive answer to the dynamics of IIR activity, this study instead raises a number of questions calling for future investigation: What affects the breakdown of users' information needs? Is it possible to predict? How do users schedule their information needs? When are users satisfied with the information collection for a subneed? Can users' satisfaction with a subneed be inferred or observed without obtrusively interrupting them? Do these propositions hold in multisession information-seeking activity? As a preliminary investigation, our study suggests some fruitful directions for future research.

### System Implications

This study suggests that because users' IIR behavior is planned, algorithms could be used to predict the plans. One promising technique is to regard plans and their manifested behavior as a hidden Markov process. If users' prior searches for the same topic have been accumulated, then plans can be inferred, and future user behavior can be predicted. The understanding of users' planning processes provides a new direction for query expansion algorithms, which are historically based on *past* behavior or corpus characteristics rather than on the anticipation of *future* need. Although planning of information need might be more difficult to predict for unstructured tasks, it is more feasible for special-purpose tasks, such as online shopping, travel planning, and online legal services.

In addition to suggesting an anticipatory approach to query revision, this study also suggests better ways to utilize past document evaluations. Proposition 2 indicates that, depending on whether the current subneed is satisfied or not, users might either employ further deepening on a subtopic, or switch to a new one. Therefore, past documents might be the best source for query expansion only when the next step is to deepen on the current subtopic, but not to switch to a new one. This study also suggests that the topicality and novelty dimensions of past documents have different uses for query expansion. Therefore, in relevance feedback, it is important to incorporate users' topicality-preserving and novelty-seeking intention and use the two dimensions selectively.

Finally, this study indicates that the standard precision-recall testing sets like those used in the Text Retrieval Conference are inadequate for the study of interactive information retrieval behavior and system performance. If system improvement is the goal, this framework calls for testing sets which record users' activity histories.

## Acknowledgment

This research is supported by the School of Computing, National University of Singapore, Research Grant: R253-000-028-112.

## References

- Bates, M.J. (1990). Where should the person stop and the information search interface start? *Information Processing & Management*, 26(5), 575-591.
- Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 1-16.
- Efthimiadis, E.N. (2000). Interactive query expansion: A user-based evaluation in relevance feedback environment. *Journal of the American Society for Information Science*, 51(11), 989-1003.
- Engeström, Y. (1987). *Learning by expanding: An activity-theoretical approach to developmental research*. Helsinki, Finland: Orienta-Konsultit.
- Fitzgerald, M.A., & Galloway, C. (2001). Relevance judging, evaluation, and decision making in virtual library: A descriptive study. *Journal of the American Society for Information Science and Technology*, 52(12), 989-1010.
- Harter, S.P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9), 602-615.
- Hiemstra, D. (1998, September). A linguistically motivated probabilistic model of information retrieval. Paper presented at the European Conference on Research and Advanced Technology for Digital Libraries (ECDL), Heraklion, Crete, Greece.
- Ingwersen, P. (1992). *Information retrieval interaction*. London: Taylor Graham Publishing.
- Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1), 3-50.
- Kekäläinen, J., & Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13), 1120-1129.
- Kirk, R.E. (1995). *Experiment design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing.
- Kuhlthau, C. (1993). *Seeking meaning: A process approach to library and information services*. Norwood, NJ: Ablex Publishing.
- Leont'ev, A.N. (1978). *Activity, consciousness, and personality*. Englewood Cliffs, NJ: Prentice-Hall.
- Lin, S.J., & Belkin, N.J. (2005). Validation of a model of information seeking over multiple search sessions. *Journal of the American Society for Information Science and Technology*, 56(4), 393-415.
- Ponte, J.M., & Croft, W.B. (1998). A language modeling approach to information retrieval. In *ACM SIGIR Conference on Information Retrieval* (pp. 275-281). New York: ACM Press.
- Tang, R., Shaw Jr., W.M., & Vevea, J.L. (1999). Towards the identification of the optimal number of relevance categories. *Journal of the American Society for Information Science*, 50, 254-264.
- Vakkari, P. (2001). A theory of the task-based information retrieval process: A summary and generalisation of a longitudinal study. *Journal of Documentation*, 57(1), 44-60.
- Vakkari, P., Pennanen, M., & Serola, S. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing & Management*, 39, 445-463.
- Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Xie, H. (2000). Shifts of interactive intentions and information-seeking strategies in interactive information retrieval. *Journal of the American Society for Information Science*, 51(9), 841-857.
- Xu, Y. (2007). The dynamics of interactive information retrieval behavior, Part I: An activity theory perspective. *Journal of the American Society for Information Science and Technology*, 58(7), 958-970.
- Xu, Y., & Chen, Z. (2006). Relevance judgment—What do information consumers consider beyond topicality? *Journal of the American Society for Information Science and Technology*, 57(7), 961-973.

## Appendix

### The Language Model Used

Hiemstra (1998) proposed a simple method to build a language model by inheriting the well-known *tf-idf* idea. When merging 10 documents from the previous round of retrieval, we extended Hiemstra's (1998) model to calculate the term frequency and document frequency for the merged document. We also defined the weight of individual documents in the merged document to be the normalized score of topicality or novelty (i.e., the 0–7 scores are converted to 0–1). Therefore, for the novelty profile, which is the merged document based on novelty scores, we have:

$$tf(w, D_N) = \sum_{i=1}^{10} n_i \cdot tf(w, D_i),$$

$$\sum_{w' \in D_N} tf(w', D_N) = \sum_{w' \in D_N} \sum_{i=1}^{10} n_i \cdot tf(w', D_i).$$

where  $tf(w, D_N)$  is the term frequency of term  $w$  in novelty profile  $D_N$ ,  $n_i$  is the novelty score normalized, and  $D_i$  refers to one of the 10 documents.  $\sum tf(w', D_N)$  is the total count of term frequency in  $D_N$ . Based on the language model (Ponte & Croft, 1998), the probability of a query term  $w$  being

produced by novelty profile  $D_N$  is defined as:

$$P_{doc}^N(w|D_N) = \frac{tf(w, D_N)}{\sum_{w' \in D_N} tf(w', D_N)}.$$

To avoid the situation in which a query term does not occur in the document and has to assume 0 as probability, a smoothing element is typically introduced in the language model. Based on (Hiemstra, 1998), the smoothing element is defined as:

$$P_{collection}^N(w) = \frac{df(w)}{\sum_{w' \in D_N} df(w')},$$

where  $df(w)$  is the frequency of documents that contain term  $w$ , and the denominator is the total document frequency for all terms in  $D_N$ . With these elements defined, the final probability of a term being produced by novelty profile is:

$$P^N(w) = \lambda P_{doc}^N(w|D_N) + (1 - \lambda) P_{collection}^N(w),$$

where  $\lambda$  is the smoothing factor and set to 0.9. Similarly, we can calculate the probability of a term being produced by topicality profile.