

Novelty and Topicality in Information Retrieval

Yunjie (Calvin) Xu*

Department of Information Systems,

School of Computing, National University of Singapore

3 Science Drive 2, 117543

Singapore

Tel: (65) 6874 6562

Fax: (65) 6779 4580

xuyj@comp.nus.edu.sg

Hainan Yin

Department of Information Systems,

School of Computing, National University of Singapore

3 Science Drive 2, 117543

Singapore

Tel: (65) 6874 6562

Fax: (65) 6779 4580

yinhaina@comp.nus.edu.sg

* Yunjie XU is the corresponding author for this paper.

This paper was accepted by *Journal of the American Society for Information Science and*

Technology in May 2007

Novelty and Topicality in Interactive Information Retrieval

Abstract

The information science research community is characterized by a paradigm split, with a system-centered cluster working on information retrieval (IR) algorithms and a user-centered cluster working on user behavior. The two clusters rarely leverage each other's insight and strength. One major suggestion from user-centered studies is to treat the relevance judgment of documents as a subjective, multidimensional, and dynamic concept, rather than treating it as objective and based on topicality only. This study explores the possibility to enhance users' topicality-based relevance judgment with subjective novelty judgment in interactive information retrieval. A set of systems are developed which differ in the way the novelty judgment is incorporated. In particular, this study compares systems which assume users' novelty judgment is directed to a certain sub-topic area and those which assume undirected. This study also compares systems which assume users judge a document based on topicality first and then novelty in a stepwise, non-compensatory fashion and those which assume users consider topicality and novelty simultaneously and as compensatory to each other. The user study shows that systems assuming directed novelty in general have higher relevance precision, but systems assuming a stepwise judgment process and systems assuming a compensatory judgment process are not significantly different.

Keywords: Relevance, relevance judgment, topicality, novelty, compensatory relevance judgment, stepwise relevance judgment, directed novelty.

Novelty and Topicality in Interactive Information Retrieval

Introduction

The information science research community is characterized by a paradigm split (Saracevic, 1999), with a system-centered cluster working on information retrieval (IR) algorithms and a user-centered cluster working on user behavior. There is a lack of integration between the two clusters. Noting the inadequate synergy between the two sides, Saracevic (1999, p.1057) lamented: “Unfortunately, in most human-centered research, beyond suggestions, concrete design solutions were not delivered. On the other hand, the system side, by and large, ignores the human side and user studies, and is even often completely ignorant of them.” Indeed, bridging and enabling synergy between the user-centered approach and the system-centered approach remain challenges to the information science research community (Ingwersen & Järvelin, 2005).

One major suggestion from user-centered researchers is that users’ relevance judgment of a document is a subjective, multidimensional, and dynamic concept, while objective measures such as the keyword match and cosine score in the vector space model could fall short of the richness of the underlying psychological processes (Borlund, 2003; Cosijn & Ingwersen, 2000). It has been repeatedly found that users’ relevance judgment encompasses not only topical match between an information need and a document, but also novelty, understandability, reliability and scope of the document (Bateman, 1998; Fitzgerald & Galloway, 2001; Wang & Soergel, 1998; Xu & Chen, 2006). Among all these criteria, novelty judgments have received particular interest by system-centered researchers (Allan et al., 2003; Brants et al., 2003; Kumaran & Allan, 2003; Yang et al., 2002; Zhang et al., 2002). Here, topicality refers to the subjective judgment of whether a document is related to the subject area of the user’s information need (Xu & Chen, 2006), and novelty refers to the

degree to which the content of a document is new to the user and beyond what the user has known before (Xu & Chen, 2006). If the relevance criteria uncovered by user studies are indeed important, the next question is how to incorporate these criteria into system design. This question is important because the actual performance of systems following the suggestions of user-centered studies offers a way to verify the validity of findings from user studies; it also provides system-centered researchers with a potential new direction to explore for innovative systems that could be better tailored to users' psychology.

The purpose of this study is to model users' relevance judgment in IR algorithm design. Particularly, this study focuses on topicality and novelty because these two dimensions are fundamental to a relevance judgment (Xu & Chen, 2006) and has attracted most interest in both user and system studies. In an interactive IR context where users' topicality and novelty judgments of documents are assumed to be available to the IR system, the aim of this study is to evaluate different ways the two types of judgment criteria could be integrated to induce a final relevance evaluation of a document. In particular, there are two possibilities: Users could consider topicality first and then novelty in a stepwise fashion, or users could consider topicality and novelty as compensatory to each other (i.e., the lack of topicality can be compensated by higher novelty, or vice versa) and judge them simultaneously. Also, there are two possibilities regarding what kind of novel documents users want: Users might regard novel documents as those having minimum redundancy with previously read documents, or users might want to continue searching for documents related to a topic area previously found novel. In other words, novelty could be assumed to be undirected or directed. We expect that if an IR system can better approximate users' relevance judgment, that is, topicality and novelty judgments, it should produce better retrieval results. A set of systems were developed based on the combination of the above possibilities. Our user test finds that systems assuming directed novelty perform better than

those assume undirected novelty, but there is an insignificant difference between systems adopting a stepwise and those adopting a compensatory decision process.

The paper is organized as follows. First, we review related user and system studies to derive our hypotheses. We then introduce a set of systems that implement these propositions to various degrees. Next we report on our user test of the systems. This is followed by result analysis and discussions.

Conceptual Background

Subjective Relevance, Topicality and Novelty

At the heart of user-centered relevance is the recognition that relevance judgments are subjective, multidimensional and dynamic. Subjectivity means that relevance is personal and individual to the users. An important aspect of relevance multidimensionality is the multiplicity of criteria in a relevance judgment. For example, Bateman (1998) listed 40 criteria that affect relevance judgments, covering aspects of content topicality, document availability, novelty, currency, information quality, presentation quality, and source characteristics. While many of these criteria could well be redundant or insignificant (Barry & Schamber, 1998; Schamber, 1994), some categories of criteria are repeatedly found important. Xu and Chen (2006) summarized five content criteria from a representative list of 13 empirical user studies: topicality, novelty, understandability, reliability and scope. Topicality and novelty emerged as two most important dimensions of relevance in their empirical test.

The task of a typical IR system is to rank (i.e., judge) documents based on their similarity with a user's request for, or indication of information need which is often quantified as a set of weighted terms. We call such quantified information need a user's

profile. For example, in a typical interactive IR system, a user's profile starts with a vector that represents terms in the initial query, and the terms and their weights can be updated later with relevance feedback technique. The vector representing a snapshot of a quantified information need is a user profile. If relevance is multidimensional, then two questions must be answered: First, can a single user profile, regardless of the way it is quantified in a vector space model, a probabilistic model or a language model, effectively capture all aspects of a relevance judgment? Second, if different aspects of a relevance judgment are to be modeled with multiple user profiles, should they be modeled and updated in the same way in an interactive IR process? These questions motivate us to explore the possibility of using a multidimensional user profile, as we shall discuss shortly.

Finally, subjective relevance judgments also stress the dynamics of judgments (Harter, 1992). The basic tenet is that users' knowledge in a domain area and users' information need are constantly modified by the information item encountered (Harter, 1992). Therefore, the judgment of topicality, novelty and other criteria evolve in the information seeking and retrieval process as users 'consume' documents.

As two major dimensions of a relevance judgment, topicality and novelty are also subjective and dynamic. However, they differ in the degree of subjectivity and dynamics. Topicality measures the 'aboutness' of a document to the topic area suggested by a query. Borlund (2003) termed topic match between the query and a document 'intellectual topicality'. It is possible for different people to agree upon the topicality of a document in relation to the query in saying that "This document belongs to the type of documents labeled by the query." That is why subject indexing in typical library catalogues is possible (Bookstein, 1979). The construction of Text Retrieval Conference (TREC, trec.nist.gov) testing set to compare different IR systems is also based on the topicality judgment of a panel of judges. Therefore, the 'relevance' precision (precision in retrieving relevant documents) of

many IR systems is more biased towards topicality precision. However, topicality does not exclude users' subjective judgment. Especially in a real search session, it is up to the user to judge whether a document is on-topic or not. Such user interpreted topicality is called subjective topicality. In our study, topicality refers to subjective topicality.

Compare to topicality, novelty is more subjective and volatile. Novelty is affected by users' background knowledge (Barry, 1994; Bateman, 1998). What one regards as novel might not be novel to another, or even to oneself a day later. A novel document can cause noticeable change in users' cognition, which in turn affects their information need and relevance judgment criteria for later documents (Harter, 1992; Zhang et al., 2002). Therefore, novelty has to be individual and dynamic. It is impossible to impose a novelty standard on a document with a majority rule. Although novelty is dynamic, the speed of change may vary depending on the importance of the search task and user's effort on learning. Therefore, learning and learning rate in a topic area are intrinsic components of a novelty judgment.

Integrating Topicality and Novelty in Relevance Judgment

To design a better IR system that integrates topicality and novelty, we shall first analyze the nature of relevance judgments. First, are relevance criteria compensatory, that is, can higher novelty compensate for lower topicality or vice versa? There are compelling arguments in information science that topicality is the first and the most important criterion for a relevance judgment. Presence of topicality as a condition for other criteria to operate is widely accepted among researchers (e.g., Schamber, 1994; Cosijn & Ingwersen, 2000; Park, 1997). Froehlich (1994, p.129) highlighted: "all relevance judgments *start* with topically relevant materials (which is an appropriate first step of system), but then diverse criteria come into play". Mizzaro (1997), in summarizing the history of relevance research, noted that relevance criteria are identified *beyond* topicality. If topicality is a necessary condition for other criteria to operate, then we should predict that if a document is off-topic, all other factors should not

matter to the relevance judgment. Therefore, relevance judgment is not compensatory when topicality is below a certain point. In that circumstance, users follow an *elimination-by-topicality* heuristic in the first step (Wang & Soergel, 1998; Greisdorf, 2003).

What if a document is judged on-topic? A user may shift her focus from topicality to the next most important attribute, say, novelty. On-topic documents are then sorted by novelty, and the best a few are accepted. This line of reasoning was clearly articulated by Boyce (1982) with a proposal for a two-stage retrieval process. In the first stage, documents are filtered by topicality. In the second stage, documents are sorted by ‘informativeness’. The resort to other criteria beyond topicality implies that topicality can be treated as a binary variable. Greisdorf (2003) depicted a stepwise judgment process starting with topicality, followed by understandability and pragmatic usefulness. Judgment at each step is regarded as binary. Therefore, Greisdorf (2003) alluded to an elimination-by-aspect process throughout the whole decision process. If the binary nature of later judgments is relaxed by allowing understandability and usefulness to be continuous, then Greisdorf (2003) would arrive at the same conclusion as Boyce’s (1982) – that the second stage can be sorted by a criterion such as novelty. Therefore, we have a *‘first eliminate-by-topicality then sort-by-novelty’* decision process. In summary, when topicality and novelty are considered, a stepwise and non-compensatory integration rule is a better approximation of users’ relevance judgment process than a compensatory and simultaneous one. Assume we can quantify the topicality and novelty aspects of users’ request for information, and call them the topicality profile and novelty profile respectively; documents can be matched against such profiles to produce a topicality or novelty score. We can hypothesize:

H1. IR systems that implement relevance judgment following the ‘first eliminate-by-topicality then sort-by-novelty’ decision process will have a better performance (as

measured by relevance precision judged by users) than the systems that treat topicality and novelty as compensatory and simultaneous.

While system-centered studies have also attempted to integrate novelty with topicality, the concept of redundancy was used in place of novelty (Allan et al. 2003; Brants et al., 2003; Kumaran & Allan, 2004; Yang et al., 2002; Zhai et al., 2003; Zhang et al., 2002). They defined redundancy as the amount of relevant information in the current document that is covered by relevant documents delivered previously. Novelty is defined as the opposite of redundancy. Such simplification might run into the risk of assuming users' learning of document content as complete, and users' novelty seeking as diversity seeking. In fact, it has been repeatedly found in user studies that in a search session, users' search objective as reflected in query revision are mainly to narrow down to a subtopic, to move to a related topic, to use different terms for the same topic, or to broaden the current topic (Bates, 1990; Nordlie, 1999; Efthimiadis, 2000; Vakkari et al., 2003). Put aside corrective actions (e.g., spelling error), if query revision can be regarded as an indication of looking for new information in a topic area, then users' selection of revision terms suggests that a new topic is often rooted in an old one. Therefore, novel documents are those which detail on certain subtopics, discuss a 'sibling' topic (Efthimiadis, 2000; Vakkari et al., 2003), or are related to the old topic in some other ways. An IR system should supply new documents that follow user's novelty direction. In other words, users' learning of novel information in the recent readings is incomplete to satisfy their information need and they want to find more. Therefore, we regard users' novelty judgment as directed. If an IR system is to dynamically provide information, such as in a relevance feedback system, it is better to assume users' novelty as directed. We hypothesize:

H2. In interactive IR systems that dynamically update users' novelty profile, systems that model users' novelty judgments as directed will have a better performance than those model users' novelty judgments as undirected.

If we cross-combine the assumptions of the compensational relationship between topicality and novelty and the assumption of novelty directedness, we have four quadrants of possible IR systems. The studies done in the system-centered cluster have been centered on the combination of non-compensatory relationship and undirected novelty (e.g., Zhang et al., 2002), with the exception (Zhai et al., 2003) which adopted compensatory relationship. However, no exploration has been done based on the hypothesis of directed novelty. Neither is the comparison of system performance based on different combinations.

Systems

In order to test our hypotheses, we propose four types of systems which are based on different combinations of the assumptions. The idea is that if the hypotheses are valid, systems that better emulate the stepwise relevance judgment process and directed novelty judgment should retrieve more relevant documents as perceived by users. Because we treat topicality and novelty as subjective judgments, our empirical study is a user study rather than the TREC-like algorithm testing.

Before we introduce the systems, we demarcate the boundary of our test. First, we assume users have only one search task in a search session which focuses on a fixed topic area. Second, our objective is not to optimize the algorithms, but rather to test the assumptions of novelty judgments, topicality judgments, and their relationship in relevance judgment. Therefore, we are more interested in observing systematic performance difference than in the magnitude of the difference which can be further improved by fine-tuning the

algorithms or exploring other alternatives. Finally, to capture the subjectivity and dynamics of topicality and novelty judgment, we use the manual feedback technique for all systems whereby users will manually evaluate the topicality and novelty of each document. The systems then use the user-assigned topicality and novelty scores to update users' information need in the form of topicality and novelty profiles for document ranking in the next round. Although automated feedback (e.g., through implicit relevance feedback) is the ultimate goal of system design, this study does not attempt to address it.

We use the popular vector space model as the base model for our algorithms as it has been found to be very robust across different applications (Baeza-Yates & Ribeiro-Neto, 1999; Zhang et al., 2002; Allan et al., 2003). The basic technique of the vector space model is to quantify documents and users' information need as weighted term vectors and then use them for similarity calculation.

For each user, her request for information can be described with two profiles: one for topicality and one for novelty, denoted as P^T and P^N respectively. A term vector with TFIDF¹ weighting is used to represent topicality profile. Novelty profile can be constructed in different ways, as we will explain shortly. Because the systems we are to compare differ in topicality-novelty integration and novelty directedness, we shall first explain how topicality and novelty profiles are built based on various assumptions. Their different combinations naturally lead to different IR systems.

¹ *TFIDF* is a popular scheme for weighting terms in a document. *TFIDF* weight $w(i,j)$ for term k_i in document d_j is defined as $w(i,j) = \frac{freq(i,j)}{\max_l freq(l,j)} \times \log \frac{N}{n_i}$, where $freq(i,j)$ is the frequency of term k_i in the document d_j , $\max_l freq(l,j)$ is the maximum term frequency among all terms occurring in document d_j , N is the total number of documents and n_i is the total number of documents containing k_i in a document set.

Topicality Profile

Topicality profile. A topicality profile is to approximate a user’s topicality judgment standard. A topicality profile is constructed with a term vector with the popular TFIDF weighting. A user’s topicality profile starts with the initial query.

Topicality profile updating strategy. The topicality profile updating strategy is based on Rocchio’s relevance feedback (Rocchio, 1971). Assume a user evaluates and assigns topicality scores to a set of documents, the topicality scores (not relevance scores) can be used to update the initial profile. We use only positive feedback. Therefore:

$$P_t^T = P_{t-1}^T + \frac{1}{|D_t|} \sum_{d_i \in D_t} d_i T_i$$

where t denotes the round of feedback in an interactive process; P_t^T is the estimate of the dynamic topicality profile for the search session at round t . D_t is the set of documents examined at round t , including both on-topic and off-topic documents retrieved in that round; $|D_t|$ is the number of documents in the set; d_i is a document vector in the set; and T_i is a subjective topicality score assigned by a user. In this formulation, a user’s current topicality evaluations (i.e., the second term in the above equation) are used to update the historical topicality profile (i.e., the first term) to form an updated estimate of the dynamic topicality profile P_t^T . There is no parameter setting to give different weights to the past and current documents. The differential contribution of a document to the dynamic profile is determined by the product of its term weights and the user’s subjective topicality judgment T_i (i.e., document topicality score).

Topicality evaluation. An algorithm’s topicality ranking of unseen documents is calculated with the similarity between the topicality profile and document vectors. We followed the vector space model and used the cosine score between a document vector and the topicality profile:

$$\text{sim}(d_i, P^T) = \cos(d_i, P^T)$$

The topicality profile is a shared component of all our systems. Systems differ in the way the novelty profile is defined and integrated with the topicality profile in document relevance evaluations.

Novelty Profile and Relevance Judgment

Novelty profile type I. A novelty profile is to approximate a user's novelty judgment standard. There are two types of novelty profiles based on different assumptions of novelty directedness. Type one assumes that users are diversity seeking. Therefore, novelty is the opposite of redundancy, and a novelty profile is basically a redundancy profile. This line of thinking started with Carbonell et al. (1998), and was followed by many system-centered researchers (Zhang et al., 2002; Allan et al., 2003; Brants et al., 2003; Yang, et al., 2002; Zhai et al., 2003), although different profile construction methods have been employed.

Carbonell et al. (1998) proposed the maximum marginal relevance (MMR) model. In this model, the marginal relevance of a document is measured with a weighted sum of its similarity to the query (which is what we defined as topicality) and its redundancy to previously selected documents. A document with high marginal relevance mandates a certain degree of relevance to the query and minimal similarity to previously retrieved documents. We apply the idea to our context to create an undirected novelty measure and to integrate topicality and novelty in a relevance judgment. Mathematically, let D be the set of all documents seen before, d_i be a new document to be judged, the redundancy score of d_i , $Rd(d_i|D)$ be:

$$Rd(d_i|D) = \max_{d_j \in D} \text{sim}(d_i, d_j),$$

and the relevance of the document can be defined as:

$$Rel(d_i|D) = \alpha \text{sim}(d_i, P^T) - (1-\alpha)Rd(d_i|D),$$

where $sim(d_i, P^T)$ measures the topicality of the document with regard to a topicality profile P^T and $Rd(d_i|D)$ is a measure of non-novelty. In this formulation, the ‘novelty’ profile of a user is basically the collection of past retrieved documents; to update the profile is to add newly retrieved documents to the collection. Novelty judgment is the redundancy judgment. In the original MMR model, the initial query was used in the place of P^T . Our adaptation incorporates dynamics into the topicality judgment. In Xu and Chen (2006), it was found that users assign topicality and novelty roughly equal weights in subjective relevance judgments. Therefore, we set α to 0.5 for one system. We also set α to 0.6 for another system based on the result of a pilot study which showed 0.6 might give a better performance. Because the choice of weight seems arbitrary, a sensitivity analysis is conducted to gauge its impact, as reported in the Simulations and Sensitivity Analysis section. In short, the above system assumes: 1) novelty judgments are undirected, and 2) novelty and topicality judgments are compensatory and simultaneous decisions. We name it the additive MMR system (MMR-Add). According to the different values for α , we name the two versions MMR-Add5 and MMR-Add6. When documents in corpus are ranked by relevance, the titles of the top 10 are displayed to users in one page in each round.

We can further adapt the MMR model to simulate the stepwise relevance judgment, whereby documents are sorted first by topicality, then by redundancy:

$$Rel(d_i) = \begin{cases} 0, & \text{if } sim(d_i, P^T) < s^* \\ Rd(d_i|D), & \text{if } sim(d_i, P^T) > s^* \end{cases}$$

where s^* is a topicality cutoff value. We use top k documents as the on-topic documents. In this study, we set it to top 20 and regard those beyond 20 as off-topic. While the cutoff point is again arbitrary, Spink et al. (2001) found that half of searchers would look at only the first two pages of returned document list (the top 20). Since we only need to return 10 documents in each round, the cutoff point of 20 also makes sure that they are among the best matches on

topicality. Nevertheless, fine tuning of this parameter is needed in future studies. In short, this system essentially assumes that for documents beyond top 20, they are treated as off-topic; for those within top 20, their relevance is based on their redundancy score. This system assumes: 1) undirected novelty judgment, and 2) stepwise relevance judgment. We name it the stepwise MMR system (MMR-Step).

Novelty profile type II. The novelty profile type II assumes directed novelty judgment. A user wants to read more on a novel aspect of the topic area until she is satisfied. Therefore, what is regarded as novel in the previous round of evaluation should be a judgment standard for the next round. In other words, the novelty judgment should be based on similarity to prior novel documents rather than the dissimilarity. We can use a term vector to represent the novelty profile P^N as we did for topicality.

TFIDF term weighting might not be the best method for a novelty profile because novelty is to be differentiated among on-topic documents while TFIDF weighting is better at differentiating different topics. Assume we have a set of novel documents and a set of non-novel documents. One way to identify novelty feature terms is to use a classification algorithm. The terms that best classify novel and non-novel documents are feature terms. To that end, we use the probabilistic measure F4 proposed by Robertson and Spark-Jones (1976). The F4 measure of a term is the ratio of relevance odds and non-relevance odds, that is, the ratio of the odds that a relevant document contains term t_j and the odds that an irrelevant document contains it. The F4 measure can be regarded as a classification measure because it assigns a weight to a term based on its relative probability in relevant and irrelevant documents. Applying the F4 measure to a set of novel and non-novel (not relevant and non-relevant) documents, the weight of t_j is:

$$w_{t_j} = \log \frac{P(t_j | N)}{1 - P(t_j | N)} - \log \frac{P(t_j | \bar{N})}{1 - P(t_j | \bar{N})} = \log \frac{r_j / (R - r_j)}{(n_j - r_j) / (S - n_j - R + r_j)}$$

where $P(t_j|N)$ is the probability that novel documents contain t_j , $P(t_j|\bar{N})$ is the probability that non-novel documents contain t_j , r_j is the count of novel documents containing term t_j , R is the count of all novel documents, n_j is the count of all documents containing t_j , and S is the total number of documents in the set. The above formulation is based on the binary classification of document novelty. To adapt it to more discrete evaluation levels, a partially novel document can contribute to both the novel and the non-novel sets proportionally. For example, if the maximum novelty score is 7, then a document with a novelty score of 5 contributes 5/7 document to the novel document set and 2/7 document to the non-novel document set. After such a conversion, r_j is the sum of novelty ‘fractions’ of only those documents containing term t_j , R is the sum of novelty scores for all documents regardless of the terms contained, n_j is the number of documents containing term t_j , and S is the total number of documents in the set.

Our pilot study showed that F4 could be slightly more effective than TFIDF. Nevertheless, as with other unigram-based algorithms, our use of the F4 weight method does not consider novelty judgment resulting from the combination of multiple terms. Such improvement should be explored in future research.

In an interactive retrieval system, documents come in rounds. Similar to traditional relevance feedback, only a small set of documents is evaluated in each round (e.g., 10). A local novelty profile can be generated for each round to capture the current judgment standard. This local novelty profile can be used to update a global novelty profile and make it dynamic. We term the local novelty profile for each round $P_{D_t}^N$, that is, the profile based on a small set of documents D_t at round t . There is a risk in calculating $P_{D_t}^N$ when all the documents are evaluated as equally novel, in which case no term has discriminating power. To avoid this situation, after collecting users’ feedback on a set of documents (e.g., the top 10), we automatically assume the bottom a few documents to be non-novel. This is inspired by the

Rocchio's (1971) suggestion on relevance feedback which assumes bottom a few documents as irrelevant. It reduces the risk of detecting no novelty terms.

The F4 measure identifies novelty feature terms regardless of the topicality implication of these terms. In the extreme case, if a stop word appears in novel documents only, it might be regarded as a good feature term. However, novelty is meaningful only when topicality exists. Therefore, for all novelty terms so identified, we multiply them with the corresponding term weights in the topicality profile. We then have a topicality-conditioned novelty profile where terms are weighed by:

$$w_{t_j} = w_{t_j}^{F4} \times w_{t_j}^{P^T}$$

In this equation, off-topic terms are discounted because of their low weights in topicality profile. This makes novelty profile a within-topic discriminator.

Updating strategy. The initial novelty profile is built based on the first round of user novelty feedback. After that, it can be updated in the succeeding rounds with the following formula:

$$P_t^N = (1 - \beta)P_{t-1}^N + \beta P_{D_t}^N$$

where $P_{D_t}^N$ is the feature term vector based on round t evaluations, and P_t^N is the global novelty profile after round t . β is an updating parameter which indicates the degree users want to stick to the novel aspects as discovered in the previous round. We set β to 0.8, which means that novelty profile is largely based on the previous round of feedback rather than rounds before. The high weight assigned to local novelty profile is base on our assumption that novelty is directed. Again, sensitivity analysis is conducted to gauge the impact of this parameter.

Novelty and relevance evaluation. With a novelty profile, a document's novelty can be evaluated with a simple cosine score between the document vector and the profile by a retrieval system. Since we can evaluate a document's topicality and novelty, two combination

strategies can be employed, with one being additive, the other stepwise. With the additive strategy, the relevance score:

$$Rel(d_i) = \gamma sim(d_i, P^T) + (1-\gamma)sim(d_i, P^N)$$

where γ is the relative weight of topicality and novelty. For the same reason mentioned above, we set it to 0.5. We call this system the additive system with directed novelty (DN-Add).

Similarly, we can define relevance as:

$$Rel(d_i) = \begin{cases} 0, & \text{if } sim(d_i, P^T) < s^* \\ sim(d_i, P^N), & \text{if } sim(d_i, P^T) > s^* \end{cases}$$

where s^* is a topicality cutoff value as before. We call this system the stepwise system with directed novelty (DN-Step).

MMR-Add5, MMR-Add6, MMR-Step, DN-Add, and DN-Step are designed based on different combinations of novelty directedness and relevance criteria integration rule. Collectively they form the set of experiment treatments for a two-factor factorial design. Comparisons of these systems' performance provide a way to test our hypotheses.

Experiment Design

Our experiment used a single search task to control the variance in system evaluation brought in by variance in search tasks (Borlund, 2000). Users were randomly assigned to one of the five IR systems outlined above. The subjects were 133 undergraduates from a major university. Our experiment was held in a computer laboratory. The experiment was carried out in multiple sessions, with about 20 students in each session. They were paid \$12 for their participation. To encourage serious participation and personal involvement, we also offered an additional award of \$50 to the one who learned the most knowledge from the search (measured with an online quiz after the search) in each session. The purpose of the additional

incentive was to motivate users to devote effort in learning the domain knowledge, which made the novelty judgment personally relevant.

Search Task

The search topic was “mobile phone radiation and health”. This was a valid topic because mobile phone has been widely accepted by the population at the experiment locale. In fact, 91% of the population at the experiment location subscribes to mobile phone services, and the mobile phone health problem was a topic of interest to many of our subjects. In the instructions given to the subjects, we described the task as follows:

Assume you are taking a health education class and the final examination which accounts for 50% of the total grade is to **search** and **study** online documents on “**mobile phone radiation and health**”. The relevance of a document depends on how much it addresses the following issues:

- *Does the use of the mobile phone pose radiation threats to the user’s health?*
- *Why is there such or no such radiation threat to health?*
- *What is the proper way to use the mobile phone to protect your health from radiation?*

You need to search for documents with the provided search engine. After a list of documents (**60 documents in six pages**) is returned by the search engine, please read each document in order, and evaluate the document in terms of whether it is **on-topic**, **novel (provides new knowledge)** to you, and **overall useful**. **You will also be asked to take a short online examination on the topic of “mobile phone radiation and health” after the search.**

However, in the search process, the search query was pre-specified in the query input box and the subjects were told that they did not need to revise the search query (the input box was not editable). The search query was “mobile phone health”. We kept the initial query the same in order to reduce system performance variance arising from different initial queries, so that we could more clearly observe the impact of search algorithms on system performance. We told the subjects that their task was only to evaluate the document returned in terms of topicality, novelty, and usefulness.

Testing Corpus

We collected the corpus for the test from Google using the following sets of keywords separately: ‘mobile phone health’, ‘mobile phone radiation’, ‘mobile phone safety’, ‘mobile phone safety precaution’, and ‘mobile phone health proper use’. We first downloaded all documents in the first 20 pages returned with these queries. We then removed duplicates and navigation pages. This gave us 295 documents (articles) including both relevant and irrelevant ones. Examining documents beyond the first 20 pages, we found only more duplicates. Therefore, we did not further expand the corpus. Also, we assumed that 295 documents should be enough to satisfy users’ information need for the topic that we had set. Most documents were published after 2001 at websites of news services (e.g., BBC News), telecommunications companies (e.g., Nokia.com), public health knowledge (e.g., myDR.com.au), and government agencies (e.g., U.S. Food and Drug Administration).

We pre-processed the collected documents by removing all html tags and irrelevant text such as header, menu, footer, ads, etc. We used only the main text of a document for indexing. Terms were not stemmed; and we obtained terms’ inverse document frequencies from a dictionary created by Berkeley University (Berkeley, 2001), which is based on a corpus of 49,602,191 web pages.

Experiment Procedure

First, we introduced the search task to the subjects. Then, on a paper survey form, they were asked to evaluate their knowledge on the topic with three subjective questions on seven-point Likert scales. Next, our research assistant explained the definition of topicality, novelty and relevance in layman terms. Relevance was explained as usefulness as suggested by past studies (e.g., Fitzgerald & Galloway, 2001). Specifically, we stated in the experiment instruction:

A document is **on-topic** if it talks about something related to your information need.

However, an on-topic document can have as little or as much content related to your information need. A document is **novel** if it provides **new** knowledge to you. A document is **overall useful** if it makes a major contribution to your information need, you expect it to substantially contribute to your quiz grade, and you try to memorize its content.

All these attributes were evaluated in an eight-point scale, with 0 indicating totally off-topic, totally non-novel, or totally useless; 1 indicating marginal on-topic, novel or useful, 4 indicating the averages, and 7 indicating very on-topic, novel or useful (Figure 1). Anchoring text was provided only for 0, 1, 4, and 7. We used eight-point scale instead of typical seven-point scale to allow for binary decision when subjects chose 0 (Xu & Chen, 2006). Our research assistant also demonstrated the system using another topic (inappropriate vitamin intake). Because the system interface resembled typical search engines such as Google, we did not find any difficulty among the subjects in using it.

<<Insert Figure 1 about here>>

Figure 1. System interface.

The subjects were randomly assigned to a system by the server. All systems had an identical interface. The first page asked the user to log in with their email address. The

second page showed them a search text box with a button. As mentioned above, the initial search keywords were fixed. When a subject submitted the query, the top 10 documents were returned and listed in one page with only the title and evaluation boxes for the three attributes (Figure 1). The first-round retrieval result was identical for all subjects because at this stage all systems performed in the identical way before users' novelty profile was created and users' topicality profile was updated. Subjects needed to read all the documents and evaluate them before they could move on to the next page. They were also asked to read the document first before giving the scores. Although the five systems updated subjects' profiles when they clicked the 'next' button and re-queried the database, the subjects were unaware of the backend feedback model when they moved to the next page. To them, the next page was just document 11 to document 20. The previously seen documents were excluded from the later rounds. Because subjects evaluated the first round differently, and the systems were different, the documents returned since the second round were different. Notice although we pre-fixed the initial search query, it was not as limited as it seemed to be – users' feedback modified their preference and their 'query' (which was represented by their profiles) was personalized since the second round. The process continued until the subjects completed six rounds of evaluation. They were then led to the quiz page. A post-experiment survey form was distributed to obtain the subjects' perception of their knowledge level on the topic. The subjects in general took one and a half hours to three hours to finish the experiment. The average time was about two hours.

In summary, the whole experiment process controlled as many variables as we could and the only difference was the systems. This allowed us to attribute system performance difference to the underlying IR models. One complication was that we could not control the difference in subjects' background knowledge or other personality factors as in typical experimental designs, but that was largely reduced by random assignment.

Result and Analysis

Among the 133 subjects in the experiment, 40 were female, 93 male, and their average age was 21 years. The subjects were experienced search engine users (mean = 6 years). They were also experienced mobile phone users (mean = 4 years). The number of subjects for DN-Add, DN-Step, MMR-Add5, MMR-Add6 and MMR-Step was 27, 28, 26, 25 and 27 respectively and satisfactory balanced. Along a seven-point scale, the subjects had an average score of 3.26 for their self-evaluated knowledge in the topic area before the experiment. After the experiment, the self-evaluated knowledge level rose to 5.15, indicating that learning did occur during the search process.

In order to evaluate system performance, we used only precision measures and consider recall as less relevant to our research objective. Our main performance measure was relevance precision, that is, the average relevance score for all retrieved documents. To obtain more information on system performance, we defined three types of precision, namely relevance precision (Pr_R), topicality precision (Pr_T), and novelty precision (Pr_N). The traditional precision measure is defined over binary document evaluation (i.e., relevant vs. irrelevant). However, our user feedback was ordinal (0-7). We normalized user feedback into a range of 0 to 1 by dividing actual user score with the maximum score (7). With the normalized score, similar to what Kekäläinen and Järvelin (2002) proposed as accumulated precision, we defined relevance precision (Pr_R), topicality precision (Pr_T), and novelty precision (Pr_N) as:

$$Pr_R = \sum_{d_i \in D} \frac{R_{d_i}}{|D|}, \quad Pr_T = \sum_{d_i \in D} \frac{T_{d_i}}{|D|}, \quad \text{and} \quad Pr_N = \sum_{d_i \in D} \frac{N_{d_i}}{|D|}$$

where D is a set of evaluated documents, d_i is a document in the set, and R_{d_i} , T_{d_i} , and N_{d_i} are the normalized relevance, topicality and novelty scores respectively for the document.

Performance Test

The unit of analysis was each subject. For each subject, precision measures can be summarized for the whole search session (e.g., 60 documents), or for each round (every 10 documents returned). Summarizing performance for the whole search allows us to compare the different systems. Summarizing at the round level allows us to drill down to compare systems at different rounds. Figure 2 reports the raw relevance precision measures at the round level.

<<Insert Figure 2 about here>>

Figure 2. Raw relevance precision

Figure 2 indicates that the five systems varied in relevance precision substantially. We also included an ‘ideal’ precision curve. This curve was obtained as follows. For all documents which were evaluated by no less than 20 subjects, we averaged its relevance scores to form an ‘objective’ relevance score. We then sorted them in descending order. The ‘ideal’ precision curve was plotted based on the first 60 documents. We ignored documents evaluated by less than 20 subjects to ensure the ‘objectivity’ of document relevance. This curve can be regarded as a practical upper-bound of an ‘ideal’ system. Obviously, the two systems augmented with directed novelty, DN-Add and DN-Step, had superior performance and were closer to the upper-bound. That being said, it is worth pointing out that theoretically this study assumes only subject relevance and *no* ‘objective’ relevance. This comparison is to illustrate what would happen if there were ‘objective’ relevance.

Figure 3 further illustrates the distribution of those documents that were among the ‘ideal’ top 60 as retrieved by different systems. Blanks in the dotted line were filled with documents outside the ‘ideal’ top 60.

<<Insert Figure 3 about here>>

Figure 3. Distribution of top 60 documents over all positions

However, direct comparison of relevance precision is not appropriate because the systems differed in relevance precision even in the first round (the means are: MMR-Add6, 0.297; MMR-Add5, 0.296; MMR-Step, 0.322; DN-Step, 0.324; DN-Add, 0.381). Theoretically, random assignment should lead to near identical relevance precision in the first round if the sample size were large enough when the subjects were evaluating the same set of documents. The difference was largely due to the randomness of the specific samples we had. In our pilot test with a smaller sample size, we observed similar inequality in the first round but in a quite different manner. To test the significance of the difference in the first round we conducted a one-way analysis of variance (ANOVA) with the dependent variable being the average relevance precision for each system and the factor being the systems. The result indicates the difference was insignificant ($F_{5,153}=2.041$, $p=0.076$). That to some degree alleviated our concern of the differences in the first round.

Notice that the subjects’ evaluation of the first round reflected their idiosyncratic characteristics in document evaluation. For instance, if one had more background knowledge, one might assign lower scores to documents in the first round as well as in the later rounds. In that sense, besides serving as a starting point to profile updating, the round 1 evaluation reflected individual difference which would also affect later rounds. In experimental design terminology, it should be treated as a covariate (Kirk, 1995). A covariate is a ‘background’ factor whose impact on a dependent variable should be factored out before one gauges the impact of main factors (Kirk, 1995). Therefore, when we compare performance for the later

rounds (i.e., 50 documents), the impact of round 1 evaluation should be factored out to give a ‘pure’ measure of the real performance difference. Also notice that for the first round, all systems by design followed the same vector space model and no novelty profile was used. Therefore round 1 did not reflect any system difference. In short, to compare the systems, round 1 should be excluded. Our comparison of system performance covered only the later five rounds.

We first took the average relevance precision of 50 documents (round 2-6) as the dependent variable. Following the analysis of covariance (ANCOVA) method with one factor (i.e., the 5 systems) and one covariate (the average relevance precision of round 1 documents), we compared the average relevance precision of the five systems. Table 1 reports the average performance of the 5 rounds (with the influence of round 1 partialled out); it also reports the significance of pair-wise differences.

<<Insert Table 1 about here>>

The result shows that DN-Step performed better than all other systems, while MMR-Step was the worst. MMR-Add6 was better than MMR-Add5, but the difference was not significant with t-test ($p=0.21$).

In the same fashion, we used ANCOVA to compare the topicality and novelty precision. Together with relevance precision, Figure 4 reports the different system performances.

<<Insert Figure 4 about here>>

Figure 4. Relevance precision, topicality precision, and novelty precision

Figure 4 suggests that in general, systems with directed novelty (DN-Add and DN-Step) performed better in topicality. The average topicality precision for DN-Add and DN-Step was 0.502 (SD=0.020), while that of the MMR-based systems was 0.460 (SD=0.020). But novelty precision was very similar between systems with directed or undirected novelty (the means

were 3.69 vs. 3.59 respectively; the SDs were 0.020 for both). It suggests that when a system retrieves documents that users considered novel in the previous round, it boosts topicality perception in this round. This is an interesting finding – when users looked for novelty by drilling down to a subtopic or switching to a related topic, they would regard such documents as more on-topic than documents that were within the same general topic area, but unrelated to what they were reading. In other words, a topic detailing on or is related to the novelty profile in the previous round is considered more on-topic in this round. Overall, system with directed novelty produced better relevance precision. In contrast, when novelty was assumed to be undirected, novelty precision did not necessarily hurt, but topicality perception suffered as a result of query drift, as reflected in Figure 4.

In order to test the hypotheses formally, we employed analysis of covariance method for our experimental design. Notice that the four systems, DN-Add, DN-Step, MMR-Add6 and MMR-Step, represent the four possible combinations based on the two assumptions, and we can regard them as four treatments of a 2×2 factorial experimental design. We consider MMR-Add6 instead of MMR-Add5 because it was generally the better system. The 2×2 factorial design allows us to evaluate the statistical significance of the novelty directedness assumption and the judgment criteria integration rule. For the four systems, we created two dummy variables: Novelty directedness (Undirected novelty=1 for MMR-Add6 and MMR-Step, otherwise 0) and Integration (Compensatory integration=1 for DN-Add, and MMR-Add, otherwise 0). Again, the dependent variable was the average relevance precision of 50 documents. Still using round 1 evaluation as a covariate, the analysis of covariance (ANCOVA) result indicates that novelty directedness had a significant main effect on system performance ($F_{1,107}=5.0$, $p=0.027$, $\eta^2=0.047$). A main effect means that regardless of other factors (e.g., the integration rule) to enhance or reduce it, novelty directedness had a baseline impact on system performance. The data reveals that systems based on directed novelty were

in general better than those based on undirected novelty (Figure 5). In contrast, the main effect of the integration method was insignificant and the difference was in the opposite direction with no practical contribution ($F_{1,107}=1.265$, $p=0.263$, $\eta^2=0.012$). In fact, contrary to our proposition, the additive integration rule outperformed stepwise systems marginally (Figure 5). We observed a moderate interaction effect (DN-Step Mean=0.434, SD=0.011, DN-Add Mean=0.350, SD=0.011, MMR-Step Mean=0.418, SD=0.011, MMR-Add Mean=0.411, SD=0.012; $F_{1,107}=3.664$, $p=0.058$, $\eta^2=0.035$). An interaction effect means that the impact of one variable (e.g., novelty directedness) is enhanced or reduced in different conditions set by another variable (e.g., integration rule). Novelty directedness had a larger impact when a stepwise non-compensatory rule was employed. However, its contribution was dampened when the compensatory rule was employed. Round 1 evaluation had a significant impact ($F_{1,107}=71.151$, $p=0.000$, $\eta^2=0.411$).

<<Insert Figure 5 about here>>

Figure 5. Interaction effect of novelty assumption and judgment criteria integration rule

In short, the examination of main effects and interaction between novelty directedness and integration rule supports hypothesis 1 that subjects' novelty judgment should be assumed to be directed. However, the result does not support hypothesis 2 that the stepwise, non-compensatory integration rule is better than the compensatory one. We observed superiority of the non-compensatory rule only when novelty is assumed to be directed.

Finally, the systems insignificantly affected the actual quiz grades. This was because the systems were not drastically different in performance, and users' learning skill and personal involvement might have played a more significant role in quiz.

Simulations and Sensitivity Analysis

We also conducted a set of simulations to further explore properties of the five systems. The five systems employed a number of parameters which were selected with heuristics. Would the superiority of some systems be a result of parameter setting? Moreover, for the MMR-based systems, novelty profile was based on TFIDF term weighting, while for the directed novelty-based systems, novelty profile was based on F4 term weighting. Did the term weighting method contribute to the different performance? We used simulation to address these questions.

The basic procedure of simulation is to use data collected for one system to run a revised system. For example, we can use the data collected from DN-Step to run a revised system with a different term weighting method. Starting from the initial query, the corpus will then be re-ranked based on the collected relevance judgments by a user. Then, using the relevance scores for each document, the user's profile can be updated, and a new round of retrieval can be carried out. However, this type of simulation faces two obstacles. First, a new system might retrieve documents not evaluated by a user. In that case, such documents have to be excluded, leaving some rounds having less than 10 documents. The severity of such missing documents on simulation reliability is a function of the number of missing documents. Second, a new system ranks documents in a different order. Therefore, a document's novelty and topicality score could change as a result of re-ordering, if they were evaluated by a real user. We have to assume that all user judgments are not changed after re-ordering. To address the first problem, we first used the most similar system to simulate a new system. Second, we simulated only rounds 2, 3, and 4 because rounds 5 and 6 tended to have many documents missing. Despite these precautionary measures, however, the impact of order effect was still not quantifiable. Therefore, we shall take the simulation result as a tentative exploration rather than confirmatory evidence.

Novelty profile updating speed. In DN-Add and DN-Step, novelty profile $P_t^N = (1 - \beta)P_{t-1}^N + \beta P_{R_t}^N$, where β is updating speed, which reflects the importance to the previous round of novelty judgment as compared to the rounds before. In the real user test, $\beta=0.8$. We simulated $\beta=0.6$ and $\beta=1$. When $\beta=1$, novelty profile is totally determined by previous-round novelty evaluation. For round 2 to 4, the missing document judgments for DN-Add were 1.9% and 0.5% for $\beta=0.6$ and $\beta=1$ respectively; for DN-Step, they were 0.6% and 0.1% respectively. The simulation result shows that the difference in performance is nearly discernable for both DN-Add and DN-Step (Figure 6). Therefore, the contribution of novelty profile is likely to be insensitive to the change in novelty updating speed in this range.

<<Insert Figure 6 about here>>

Figure 6. Sensitivity of DN-Add and DN-Step to novelty updating speed

Novelty weight in relevance evaluation. In MMR-Add5 and MMR-Add6, the relevance of a document: $MMR(d_i|R) = \alpha sim(d_i, P^T) - (1-\alpha)Rd(d_i|R)$, where $(1-\alpha)$ is the weight for redundancy. Using data in MMR-Add5, we simulated the system when α is set to 0.6, 0.7, 0.8, 0.9 and 1. The missing document evaluations were 17.7%, 23.1%, 27.2%, 31.5% and 32.6% respectively. The result (Figure 7) indicates that 0.6 was the best performance when averaged over round 2 to 4. In agreement with this, when we regressed document relevance scores on topicality and novelty scores, the regression coefficients were 0.56 and 0.38 respectively, suggesting 0.6 is a good approximation. In the user test, MMR-Add6 also produced satisfactory result. Overall, the result suggests that the superiority of DN-Step over redundancy-based systems is not likely to be changed by fine-tuning of the parameter for redundancy-based systems.

<<Insert Figure 7 about here>>

Figure 7. Sensitivity of MMR-Add5 to redundancy weight

We also simulated weights for novelty in DN-Add, where $Rel(d_i) = \gamma sim(d_i, P^T) + (1 - \gamma) sim(d_i, P^N)$. Based on the user data collected for $\gamma=0.5$, we simulated $\gamma=0.0$, $\gamma=0.3$ and $\gamma=0.7$. There were 12.6%, 23.5% and 28.6% missing document evaluations for the three levels of γ respectively. There seemed to be sensitivity to the parameter. The result (Figure 8) indicates that the performance of DN-Add might be further improved through parameter setting. The overall performance was better when $\gamma=0.7$. Future user study is needed here. However, the conclusions regarding the hypothesis 1 is not likely to be changed. But the overall performance of stepwise, non-compensatory rule might be improved.

<<Insert Figure 8 about here>>

Figure 8. Sensitivity of DN-Add to novelty weight

Novelty profile based on TFIDF weighting. Because we used different term weighting methods for directed novelty-based systems and MMR-based systems, a complication is that whether the superiority of directed novelty-based systems was a result of different weighting methods. We simulated DN-Add and DN-Step with a profile based on TFIDF weighting. The missing document evaluations were 26.4% and 17.1% for DN-Add and DN-Step respectively. In contrast to our pilot test, the result (Figure 9) indicates that TFIDF weighting performs marginally better (3% and 1% respectively), but the difference was insignificant. Overall, the set of simulations do not undermine the validity of the conclusions we have drawn for the hypotheses.

<<Insert Figure 9 about here>>

Figure 9. Simulation of DN-Add and DN-Step with TFIDF weighting

Discussion and Conclusion

We have set out to test the hypotheses that novelty judgment is directed rather than undirected, and that the stepwise, non-compensatory integration rule is better than the simultaneous, compensatory rule in IR systems with multiple relevance criteria.

Our experimental analysis shows significant main effect for the novelty assumption, suggesting directed novelty in general approximates users' relevance judgment better. In fact, with DN-Add, and DN-step, introducing novelty profile resulted in more than improved novelty precision; topicality precision also improved, meaning the benefit of novelty profile spills over to topicality. This is likely as a directed novelty profile helps users narrow their search to a specific subtopic where they may find more on-topic information. This finding essentially suggests that a relevant content element is first perceived novel then on-topic if the user decides to pursue in that direction.

Our study reveals that the contribution of a novelty profile is dependent on how it is conceptualized and operationalized. When novelty is defined as the opposite of redundancy, even when the documents retrieved are considered novel, the topicality of documents might not improve. This concurs with the finding of Zhai et al. (2003) that novelty might come at the price of topicality. In their system, topicality and novelty were compensatory. System performance was affected when low topicality high novelty documents were ranked high. Our result suggests that even when we use non-compensatory rule (e.g., MMR-Step) to 'protect' topicality from drifting, topicality is still sacrificed if novelty is defined as redundancy. This is understandable because in both the Zhai et al.'s (2003) study and ours, when pursuing for novelty, redundancy cannot anchor documents to a desirable (sub)topic, which leads to query drift. In that sense, novelty judgment should better be considered directed rather than undirected.

Integration rule seems to make little difference to retrieval performance. However, it moderates the effect of novelty. When undirected novelty is assumed (for MMR-Add6 and MMR-Step), a non-compensatory integration rule makes a significant negative impact. A compensatory rule is better in that case. This is probably because the compensatory rule dampens the effect of query drift. In our experiment, when directed novelty was assumed (for DN-Add and DN-Step), the non-compensatory integration rule showed an insignificant positive impact. A plausible reason for the modest contribution of the non-compensatory rule in this case was that our corpus was built based on keyword search, which eliminated many off-topic documents in the first place. Therefore, the extra contribution of the second-time elimination-by-topicality was much weaker than it could be. Future study with a more natural corpus is needed to verify this proposition.

Finally, although not intentionally designed, our data analysis shows that individual difference as reflected in round 1 document evaluation is a good predictor of users' relevance judgment in the later rounds. In comparison, the different designs of a system contribute relatively less. On one hand, this implies more should be done to improve the algorithms; on the other hand, it implies that capturing subjectivity and individual difference in IR is really a profitable area of research although its potential is far from being realized.

We acknowledge a number of important limitations that may threaten the validity of our findings. First, our result is based on only one search task. The validity of our findings needs to be verified with other studies of similar setting but with different search topics. Furthermore, the artificial constraint that disallowed subjects to edit query might have made the experiment even less realistic to subjects. Second, it is beneficial to further increase the sample size of our subjects, as reflected in substantial difference in the round 1 evaluation across systems. If the impact of the round 1 evaluation on later document evaluation is nonlinear, then the conclusions that we have discussed earlier might be in question. Third, it

is desirable to test parameter sensitivity with real user groups rather than simulations. Forth, the use of a small pre-filtered corpus might introduce bias in the conclusion. Fifth, we have not explored fine-tuning of parameters in this study. Finally, it is desirable to compare system performance in realistic settings rather than in an experimental setting because users' learning behavior could be different in a real setting.

Despite its limitations, this study represents an effort toward integrating findings from user-centered studies and system-centered studies. We have explored the possibility of quantifying relevance judgment as a multidimensional construct. Particularly, we have explored the two important dimensions of relevance: topicality and novelty. We have confirmed a hypothesis drawn from user-centered studies and offer them an empirical grounding in system studies. We have also provided a set of practical implementation methods as called for by Xu and Chen (2006). Our preliminary findings open the way for better quantification methods to be proposed to describe user topicality and novelty profile in future studies.

Acknowledgement

This research is supported by the School of Computing, National University of Singapore, Research Grant: R253-000-048-112.

Reference

- Allan, J., Wade, C., & Bolivar, A. (2003). Retrieval and novelty detection at the sentence level. In *the Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 314-321), Toronto, Canada, July 28-August 1, 2003. New York, NY: ACM.
- Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*, ACM Press / Addison-Wesley.

- Barry, C., & Schamber, L. (1998). Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing & Management*, 34, 219–236.
- Barry, C. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, 45(3), 149-159.
- Bateman, J. (1998). Changes in relevance criteria: A longitudinal study. In *the Proceedings of the 61st Annual Meeting of the American Society for Information Science*, 23–32. Medford, NJ: Information Today.
- Bates, M. J. (1990). Where should the person stop and the information search interface start? *Information Processing & Management*, 26(5), 575-591.
- Berkely, University of California. (2001). Web Term Document Frequency Form. <http://elib.cs.berkeley.edu/docfreq>. Accessed on June 9, 2005.
- Bookstein, A. (1979). Relevance. *Journal of the American Society for Information Science*, 30(5), 269-273.
- Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56(1), 71-90.
- Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for information Science and Technology*, 54(10), 913-925.
- Boyce, B. (1982). Beyond topicality: A two stage view of relevance and the retrieval process. *Information Processing & Management*, 18(3), 105-109.
- Brants, T., Chen, F., & Farahat, A. (2003). A system for new event detection. In *the Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 330-337), Toronto, Canada, July 28 - August 1, 2003. New York, NY: ACM.
- Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *the Proceedings of the 21st*

- Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 335-336), Melbourne, Australia, August 24-28 1998. New York, NY: ACM.
- Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management*, 36(4), 533-550.
- Efthimiadis, E. (2000). Interactive query expansion: A user-based evaluation in relevance feedback environment. *Journal of the American Society for Information Science*, 51(11), 989-1003.
- 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.
- Froehlich, T. J. (1994). Relevance reconsidered: Towards an agenda for the 21st century: Introduction to special topic issue on relevance research. *Journal of the American Society for Information Science*, 45(3), 124 –134.
- Greisdorf, H. (2003). Relevance thresholds: A multi-stage predictive model of how users evaluate information. *Information Processing & Management*, 39, 403-423.
- Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for information Science*, 43(9), 602-615.
- Ingwersen, P., & Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*, Dordrecht, Netherlands: Springer.
- 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). *Experimental Design*. Pacific Grove, CA: Books/Cole Publishing.

- Kumaran, G. & Allan, J. (2004). Text classification and named entities for new event detection. In *the Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 297-304), Sheffield, UK, July 25-29, 2004. New York, NY: ACM.
- Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9), 810–832.
- Nordlie, R. (1999). “User revelation” – A comparison of initial queries and ensuing question development in online searching and in human reference interaction. In *the Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 11-18.), Berkley, CA. New York, NY: ACM.
- Park, H. (1997). Relevance of science information: origins and dimensions of relevance and their implications to information retrieval. *Information Processing & Management*, 33(3), 339-352.
- Robertson, S. E., & Spark-Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27, 129-146.
- Rocchio, J. J. (1971). Relevance feedback in information retrieval. In Salton, G. (ed.), *The SMART Retrieval System - Experiments in Automatic Document Processing* (Chapter 14, pp. 313-323). Englewood Cliffs, NJ: Prentice-Hall.
- Saracevic, T. (1999). Information science. *Journal of the American Society for Information Science*, 50(12), 1051-1063.
- Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology*, 29, 33-48.

- Spink, A., Wolfram, D., Jansen, M., & Saracevic, T. (2001). Search the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226–234.
- Vakkari, P., Pennanen, M., & Serola, S. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing and Management*, 39, 445-463.
- Wang, P., & Soergel, D. (1998). A cognitive model of document use during a research project. Study I. Document selection. *Journal of the American Society for Information Science*, 49(2), 115–133.
- Xu, Y., & Chen, Z. (2006). Relevance Judgment – What Do Information Users Consider beyond Topicality? *Journal of the American Society for Information Science and Technology*, 57(7), 961-973.
- Yang, Y., Zhang, J., Carbonell, J., & Jin, C. (2002). Topic-conditioned novelty detection. In *the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 688-693), Edmonton, Alberta, Canada 2002. New York, NY: ACM.
- Zhai, C., Cohen, W. W., & Lafferty, J. (2003). Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *the Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 10-17), Toronto, Canada, July 28 - August 1, 2003. New York, NY: ACM.
- Zhang, Y., Callan, J., & Minka, T. (2002). Novelty and redundancy detection in adaptive filtering. In *the Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 81-88), Tampere, Finland, August 11-15, 2002. New York, NY: ACM.

| Title | Novel | On-Topic | Useful |
|---|----------------------|----------------------|--|
| <u>Mobile phone</u> | <input type="text"/> | <input type="text"/> | <input type="text"/> |
| <u>Don't give mobile phone to the under-9s</u> | <input type="text"/> | <input type="text"/> | 0-USELESS 1-A BIT USEFUL 2 3 4-SATISFACTORY 5 6 7-ESSENTIAL |
| <u>It's a mobile world</u> | <input type="text"/> | <input type="text"/> | |
| <u>Using GSM-health & environment issues-FAQs</u> | <input type="text"/> | <input type="text"/> | |
| | | | |

Figure 1. System interface.

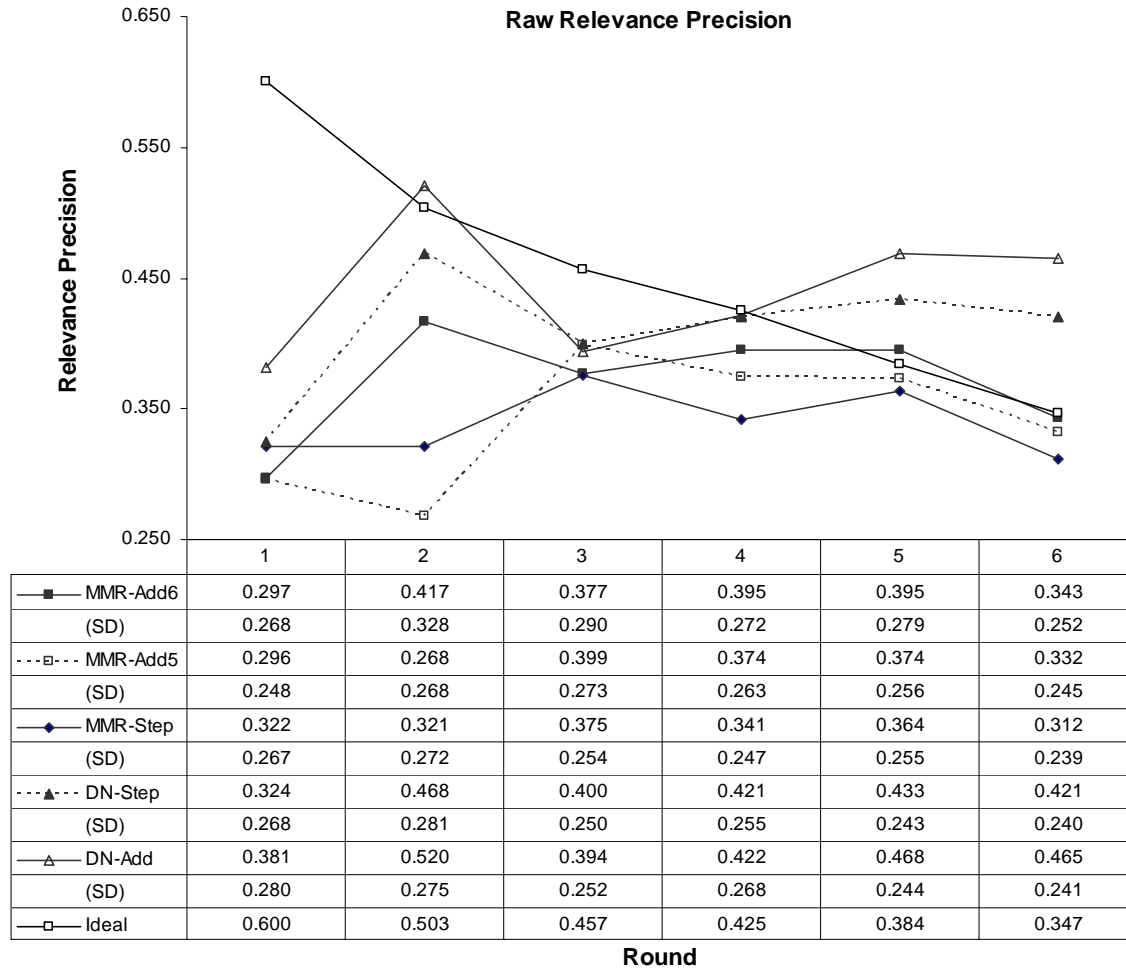


Figure 2. Raw relevance precision.

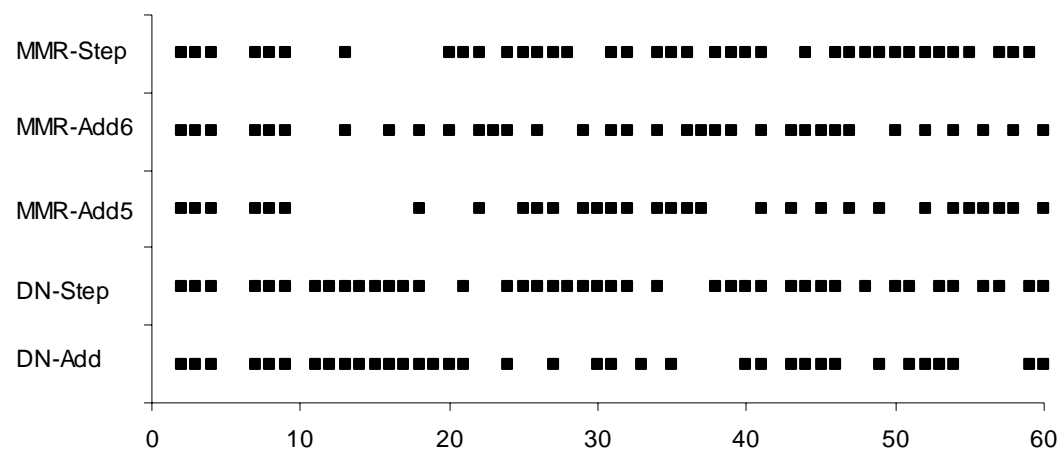


Figure 3. Distribution of the top 60 documents over 60 positions

Relevance, Topicality and Novelty Precision

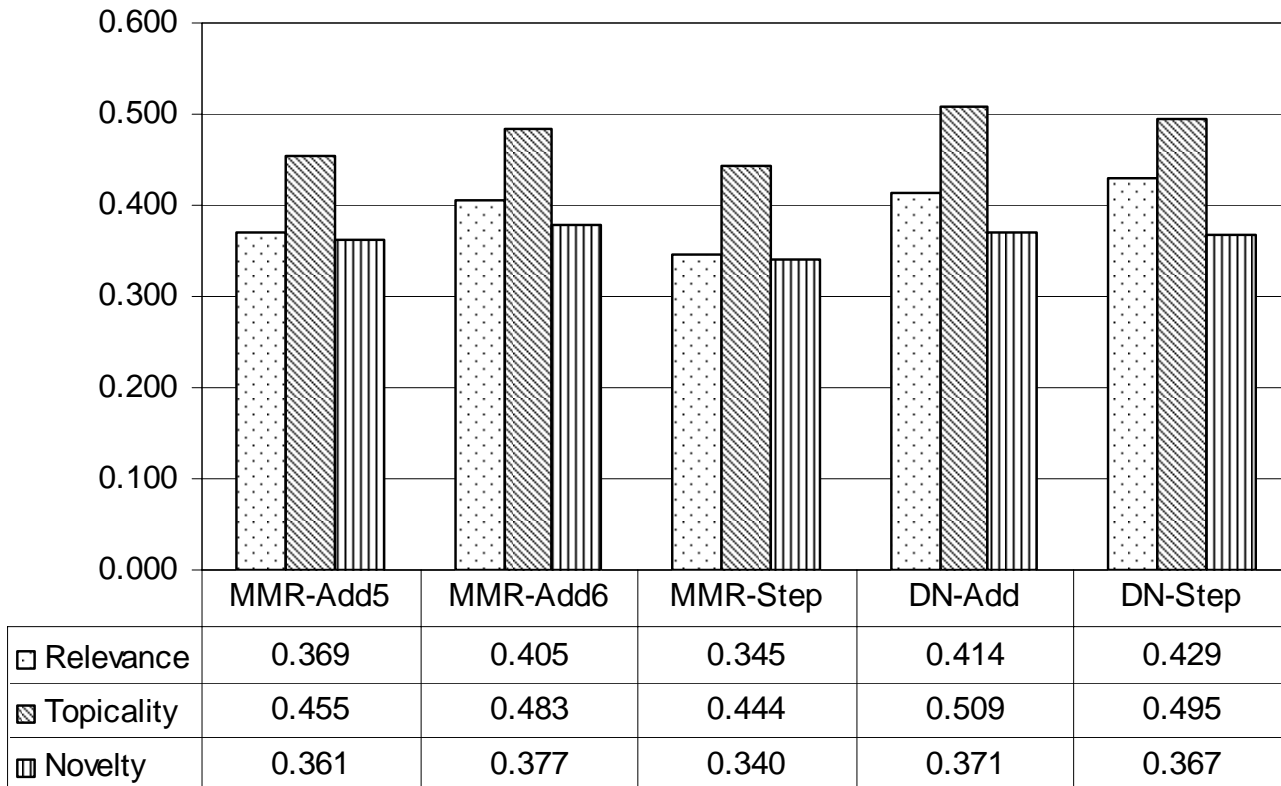


Figure 4. Relevance precision, topicality precision, and novelty precision.

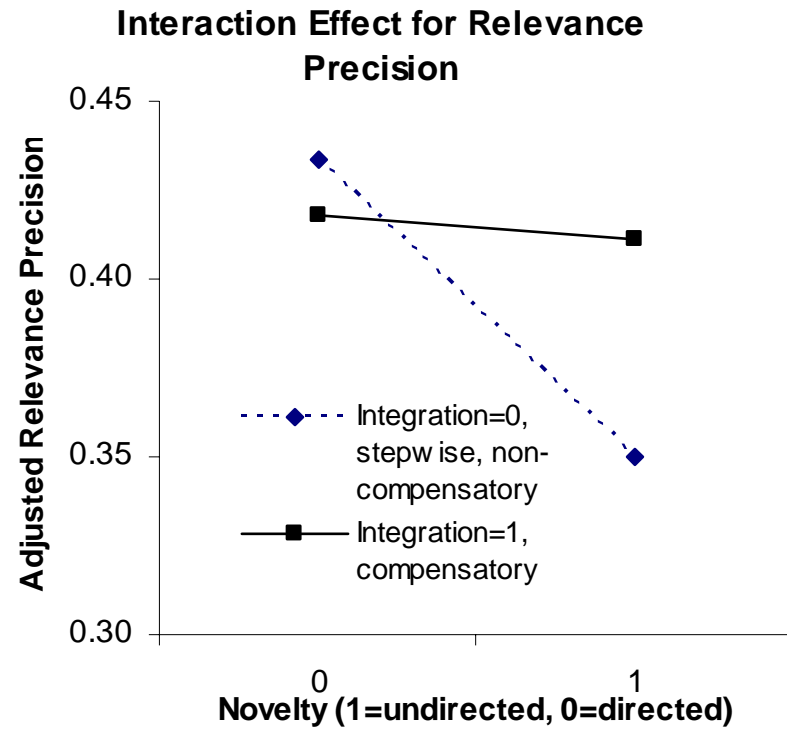


Figure 5. Interaction effect of novelty assumption and judgment criteria integration rule.

Sensitivity of DN-Add and DN-Step to Novelty Updating Speed

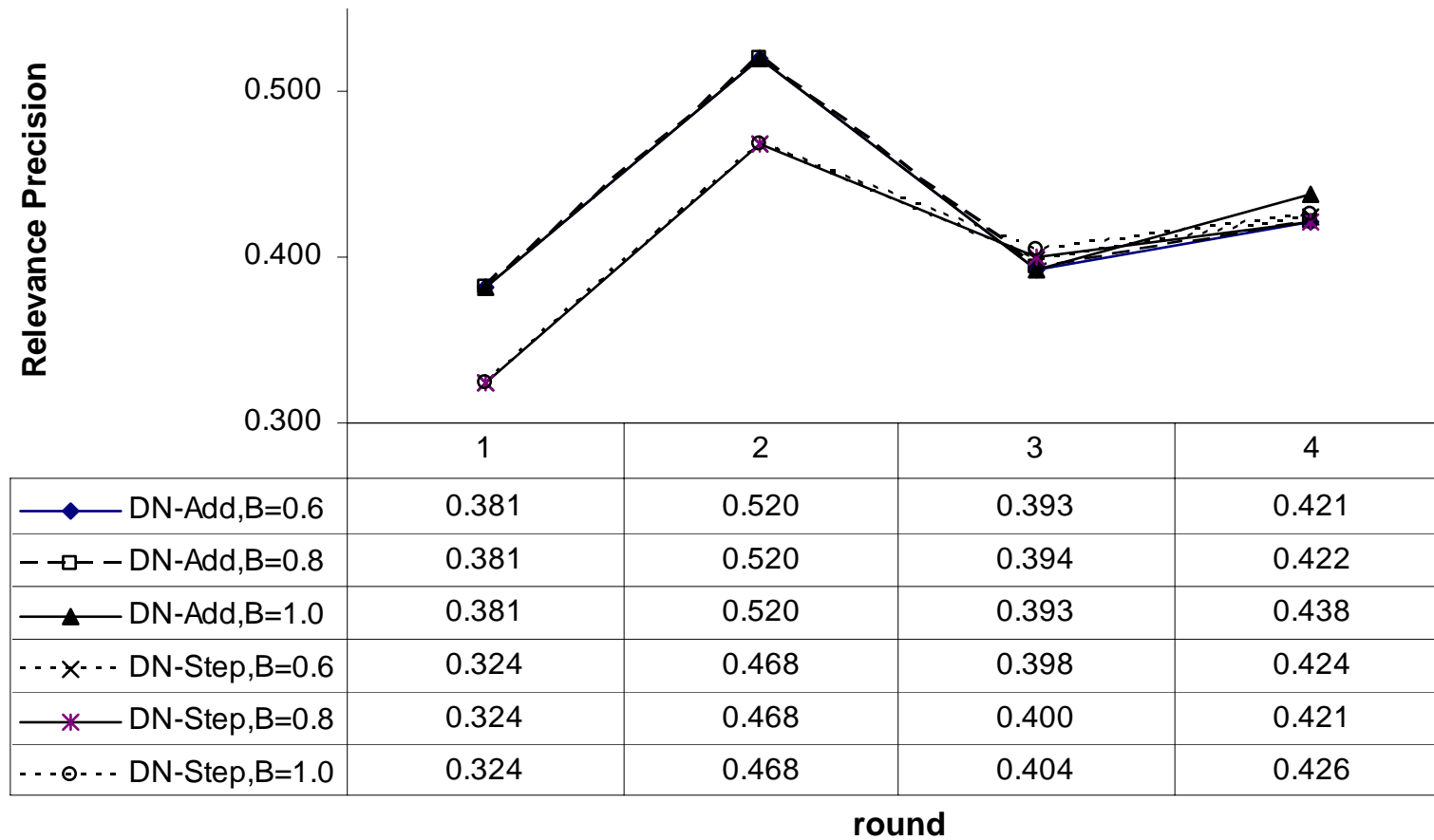


Figure 6. Sensitivity of DN-Add and DN-Step to novelty updating speed.

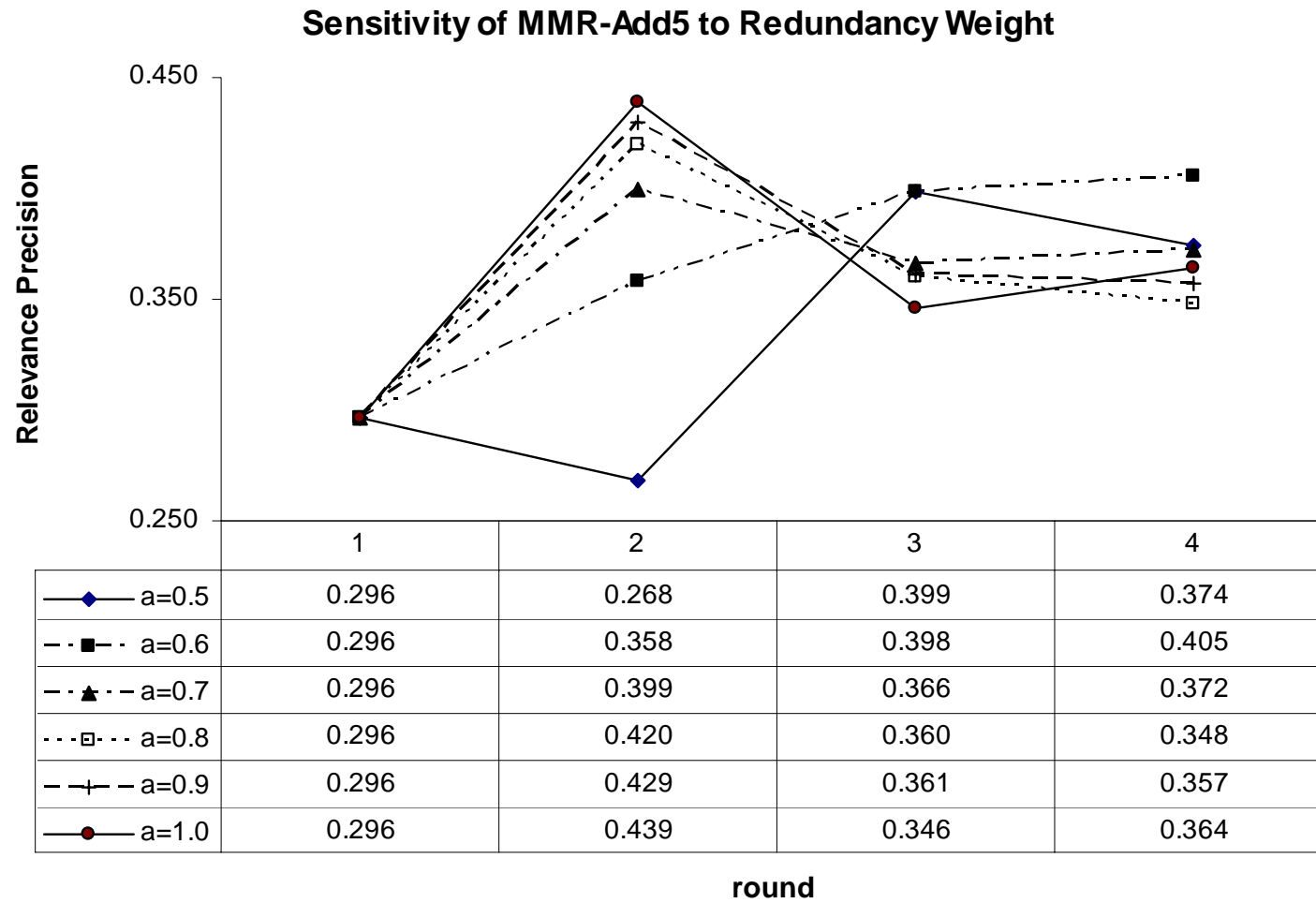


Figure 7. Sensitivity of MMR-Add5 to redundancy weight.

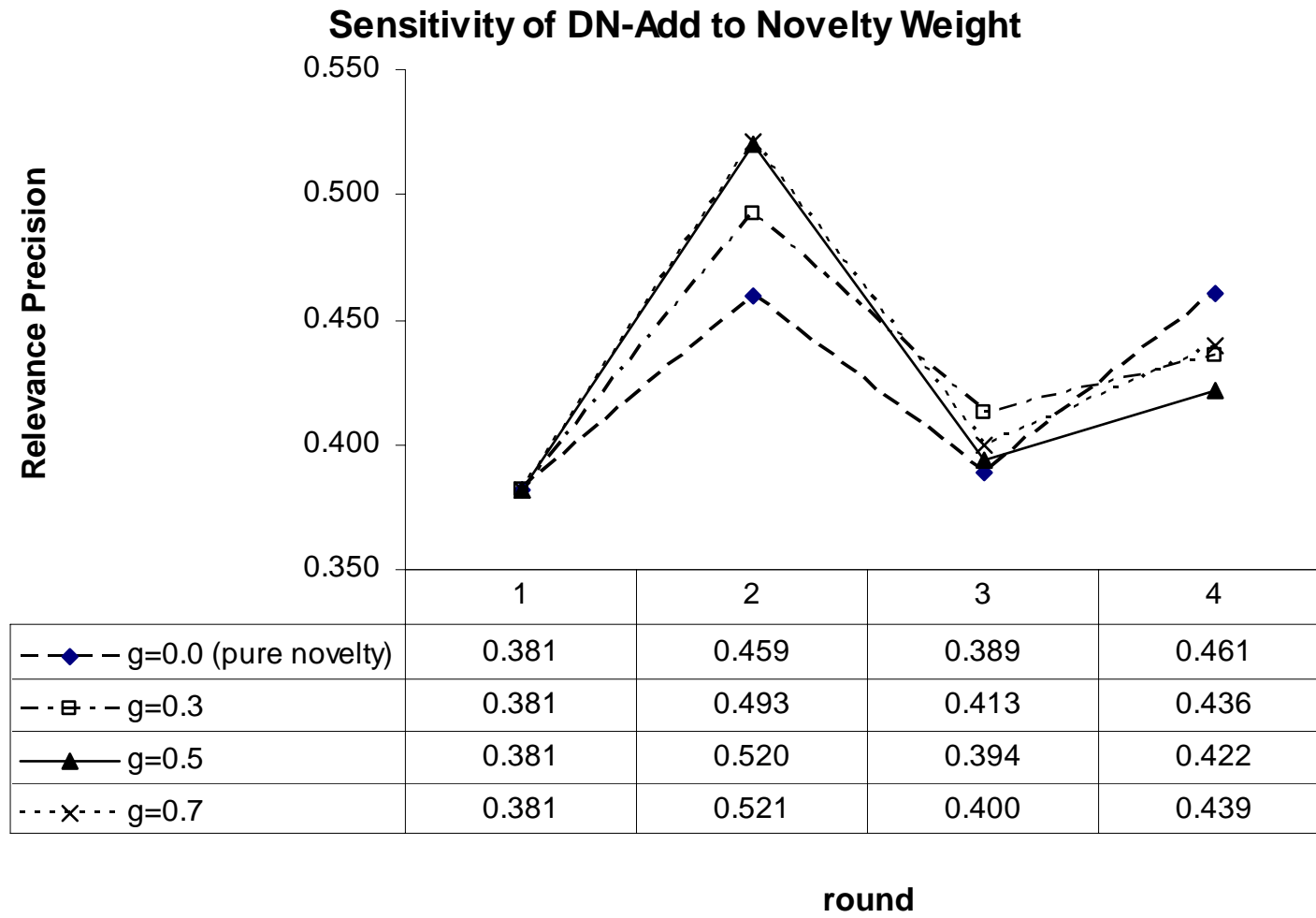


Figure 8. Sensitivity of DN-Add to novelty weight.

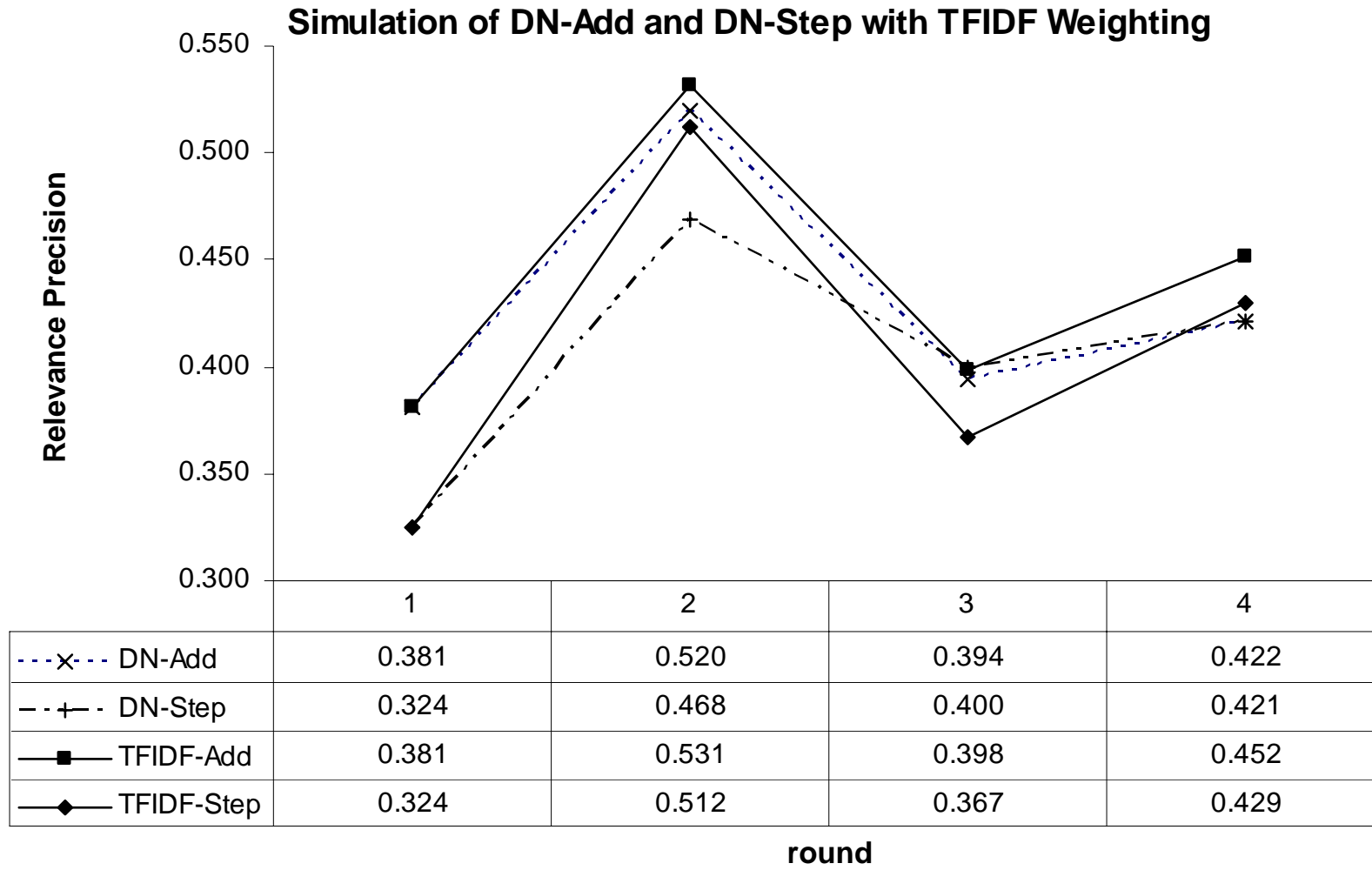


Figure 9. Simulation of DN-Add and DN-Step with TFIDF weighting.

Table 1. Relevance precision of different systems

| Systems | Mean | Std. Error | Systems - DN-Add | Systems - MMR-Add5 | Systems - MMR-Add6 | Systems - MMR-Step |
|----------|-------|------------|---------------------|-----------------------|-----------------------|-----------------------|
| DN-Add | 0.414 | 0.020 | | | | |
| MMR-Add5 | 0.369 | 0.020 | -0.044 | | | |
| MMR-Add6 | 0.405 | 0.020 | -0.009 | 0.036 | | |
| MMR-Step | 0.345 | 0.020 | -0.069* | -0.025 | -0.061* | |
| DN-Step | 0.429 | 0.019 | 0.015 | 0.059* | 0.024 | 0.084** |

* Significant at $p < 0.05$ level; ** Significant at $p < 0.01$ level