

# Collaborative Relevance Judgment: A Group Consensus Method for Evaluating User Search Performance

Xiangmin Zhang

Library and Information Science Program, Wayne State University, Detroit, MI 48202. E-mail: ae9101@wayne.edu

**Relevance judgment has traditionally been considered a personal and subjective matter. A user's search and the search result are treated as an isolated event. To consider the collaborative nature of information retrieval (IR) in a group/organization or even societal context, this article proposes a method that measures relevance based on group/peer consensus. The method can be used in IR experiments. In this method, the relevance of a document is decided by group consensus, or more specifically, by the number of users (or experiment participants) who retrieve it for the same search question. The more users who retrieve it, the more relevant the document will be considered. A user's search performance can be measured by a relevance score based on this notion. The article reports the results of an experiment using this method to compare the search performance of different types of users. Related issues with the method and future directions are also discussed.**

## Introduction

Issues in relevance and relevance judgments have been inviting researchers' attention for decades because they are central to information retrieval (IR) and the basis of nearly all experimental evaluations/testing of IR systems and user performance. There are various views towards relevance and various ways to assess relevance. Saracevic (1975) analyzes relevance comprehensively from different views in earlier years of information science. Since the late 1980s, the focus of the research on relevance has shifted from a system's perspective to the user-centered point of view. User-centered relevance research concentrates on how relevance is judged from the user's perspective and the criteria users would use (e.g., Barry, 1994; Schamber, 1994). The results of these studies reveal that relevance is a dynamic process and users judge relevance based on many criteria.

Despite the shift from the system's perspective to the user-centered research, the basic thinking on relevance and

relevance judgment has rarely changed, i.e., it is a subjective, personal process (Saracevic, 1975). The issue of subjectivity has been noted, but no solution has been proposed (Mizzaro, 1997). Based on this view, whether or not a document is relevant to the search topic and/or the searcher's information need, is completely decided by the searcher's judgment. This judgment is considered to be the searcher's psychological or cognitive process (Harter, 1992).

Although not denying the importance of a user's personal, private judgment, the very subjective nature of this relevance judgment thinking has two major limitations. First, it excludes considerations of the social interactions or collaborations of IR in real-life situations and, therefore, does not completely reflect the reality of these situations. In these situations, searches may occur in a group or organizational context. It is common for the searcher to seek help from others and to share the same opinions about the search results with peers or colleagues. Unfortunately, "IR research has been somewhat slow in recognizing the power latent in social relations that tie the users of information systems together . . ." (Karamuftuoglu, 1998, p. 1070).

Second, the subjectivity of the relevance judgment makes it inconsistent and thus hard to measure. This raises problems in determining *precision* and *recall*, two classical measures based on relevance judgment. Recall is defined as the ratio of the number of relevant documents retrieved to the number of relevant records existing in the database. Precision is defined as the ratio of the number of relevant records retrieved to the total number retrieved (Boyce, Meadow, & Kraft, 1994, pp. 180–181; Salton & McGill, 1983, Chapter 5; van Rijsbergen, 1979, Chapter 7). The two measures have frequently drawn criticism from researchers (e.g., Blair, 1996; Hersh, 1994; Meadow, Marchionini, & Cherry, 1994; Su, 1994; Yee, 1993). Recall is technically hard to measure because it requires the knowledge of all relevant documents in the database(s) to be used in the test. Precision has been found to be a poor indicator of search success (Tiamiyu & Ajiferuke, 1988; Su, 1994). Su (1994) found that precision was not significantly correlated with the user's judgment of IR success. The two measures tell little about whether the user has obtained articles that would

---

Received February 28, 2001; Revised July 27, 2001; accepted October 16, 2001

© 2002 Wiley Periodicals, Inc.

DOI: 10.1002/asi.10036

enable him or her to get the desired information (Hersh, 1994).

Subjective judgment by an individual may be a decisive factor for relevance. However, the consensus of peers or group members need to be taken into account in relevance judgments. In reality, IR is often a collaborative process, as is relevance judgment. A person's judgment is often influenced by other people's opinions. The searcher may deem a document relevant simply because it is recommended by a colleague or it is referred to in the discussion of a related issue by peers.

The present article proposes a method for assessing relevance, which includes the consideration of group or peer consensus and can be used for evaluating user performance in IR experiments. The method also defines a measure that makes relevance judgment relatively objective, without the involvement of external human assessors who are not real users. However, the proposed method is not intended to ignore the subjective nature of relevance judgment by end users, but to serve as a complement to it.

## Related Research

### *Collaboration in Information Seeking*

Collaboration in information seeking is a common practice. People often help each other in various ways to find needed information. Karamuftuoglu (1998) discusses the collaborative nature of IR from the point of view of knowledge production. The author argues that given the collective nature of knowledge production, of which information seeking is a part, the conception of users of IR systems as private individuals can be questioned. "Not only may a user be undecided to a varying degree in relation to a document at any given moment, but a number of other 'subjects' may be engaged with the same document with varying degrees of interests at any given time. Collaboration of a number of individuals with shared interests is of fundamental importance to the practice of knowledge production (e.g., research). Existence of such a community is the basic necessity of the practice of science . . ." (Karamuftuoglu, 1998, p. 1074).

Collaboration in IR is further evidenced by Twidale's research. Twidale, Nichols, and Paice (1997) have observed collaborations between users conducting information retrieval in physical library settings. The collaborations include joint search (a group of people work around a single terminal); coordinated search (a group works on two or more adjacent terminals and discuss what they are doing . . .); query (ask another person working in the area for help); chance contact (accidentally found somebody else's printout interesting and ask whose it is), etc. Collaborations in using computerized IR systems were also identified. For example, users would ask around (during a search): "Do you know? . . . Does anyone know?" etc. (p. 766).

In another study, Twidale and Nichols (1998) discuss the features of collaboration in IR. They conclude that information searching is part of people's larger work activities

that generally involve some interaction with colleagues. These interactions can include recommendations of relevant items, the sharing of search tactics, and informal explanations and help about how to use a particular information system.

### *Sharing of Information (Search Results)*

Of particular interest to the present study is the sharing of the search results. Collaborations in IR naturally lead to the sharing of information that was generated as the search results. Karamuftuoglu (1998) points out that "in the context of documentary IR, information sharing of course means sharing the search results, or better, sharing the data about the whole of a search session" (p. 1075). The search results shared may be from the same search session, or may be from a search session a long time ago. They can be shared via personal and/or group recommendation with known individuals or group members and with unknown users whose interests are similar to the information seeker. In our daily life, people ask their friends and colleagues for information related to a specific topic. The most familiar information sharing activity can be observed in USENET Newsgroups on the Internet. Postings asking for textbook recommendations for a particular topic frequently appear in newsgroups. People who are familiar with the related books will respond with recommendations. Rioux (2000) studied information sharing behavior on the Web, and made suggestions to improve electronic sharing tools.

Twidale et al. (1997) also point out that sharing search results is common. "Retrieved information may be passed on to other unknown users whose interest profiles resemble the present searcher" (p. 772).

However, in the existing IR systems, the product of a successful search is rarely made available to other users, except perhaps one or two close colleagues. The result is that many searches are essentially identical to searches that have already been done, perhaps many times over by somebody else. Users would benefit greatly if the products of past searches could be shared with other interested users (Twidale et al., 1997).

### *Sharing of Relevance Judgments*

Information sharing implies the sharing of the same or similar relevance judgments on the same piece of information. Otherwise, it will be meaningless to share that piece of information. Research in user-centered relevance judgment has indicated that consensus is one of users' relevance criteria.

In Barry's (1994) study, criteria used by real users for relevance judgment were investigated. Among the revealed criteria employed by experimental users, *consensus within the field* is found to be one of them, which is defined in the study as the extent to which there is or is not consensus within the intellectual field relating to the information within the document. The study found that whether or not there is a consensus on a particular theory or question

seemed to affect respondents' decisions to pursue the information (Barry & Schamber, 1998).

In their study on scholars' judgment of information quality and cognitive authority found on the Internet, Rieh and Belkin (2000) identified several groups of criteria for judgment used by people. One category they found was "one-collective source." They found that to some scholars, what mattered was whether the information is based on a single person's opinion or on that of group of people including an organization, a company, or an institute. Cases were found where the retrieved information was trusted because it had been referenced by multiple sources.

In addition to the explicit evidence of the group consensus criterion discussed above, the similar idea can be found implemented in some information systems. For example, in recommender systems, recommendations are provided to users by others the user may not know (Resnick & Varian, 1997). Recommendations from other people represent the experience and judgment of a community. In everyday life, it is common that people rely on recommendations from other people by various means. By sharing recommendations, a user without sufficient personal experience/knowledge can benefit from the relevance judgment based on the prior experience of others.

Hill, Stead, Rosenstein, and Furnas (1995) reported a study on a video recommender system. The system was constructed to compare a viewer's personal ratings of videos with those of hundreds of others to find people with similar preferences. The system then recommends to the viewer unseen videos that these similar people have viewed and liked. The main purpose of the study was to see if this kind of recommendation would work. People were invited to participate in the study through e-mail. When participants sent a message to the system's e-mail address, the system would reply with an alphabetical list of 500 videos for them to rate. Each person rated the titles they had seen on a scale of 1 to 10, where 1 is low and 10 is high. Users may also rate an unseen movie as "must-see" or "not interested," as appropriate. The data set included more than 55,000 ratings of 1750 movies by 291 users.

When a new person entered ratings, the database was searched to find a sample of known users with similar ratings. This new person's preferences could then be modeled as a linear combination of the rating from the best matching people. In this way the system could suggest movies that the model predicted the user would like but that the user had not rated (and presumably not seen). Feedback about the system's recommendations was invited from the users. Out of 51 responses, 32 (63%) were positive, 14 (27%) negative, and 5 (10%) neutral.

A similar system is RINGO (Shardanand & Maes, 1995) that makes personalized music recommendations to individual users. The system asks people to rate some music based on their listening pleasures. These ratings are saved in the person's profile. RINGO compares user profiles to determine which users have similar taste (like/dislike the same albums). The system can predict how much a user may like an album/artist that has not yet been rated by the user, based

on similar users' ratings. People's tastes and their ratings on a specific album/artist in this system are similar to relevance judgments on documents in IR systems. Therefore, the same idea can be used for IR systems.

Examples exist in IR systems, too. When constructing search agents for users, Newell (1997) found that instead of simply conducting a search using traditional methods, a user may obtain better search results by using an existing agent created by another user with a similar background to do a similar search. The search results generated by the existing agent would be of interest to this user.

Text REtrieval Conferences (TREC) use a pooling method to assemble relevance assessments (Harman, 1995; Voorhees, 2000). In TREC's method, which has been relatively unchanged since TREC-3 (1994), the human relevance assessors create search topics and use them to search the test document collection to find relevant documents. Participants then use the topics to make a retrieval run with their system and submit a ranking of the top 1,000 documents for each of the topics to NIST (National Institute of Standards and Technology, the sponsor for TREC conferences) for evaluation. Two or three different runs may be submitted. After all participating systems have submitted their document rankings, NIST forms a pool of possibly relevant documents for each topic from the top 100 documents from each run that will be judged. The human relevance assessor who created the topic makes the final relevance judgment for each document in the pool: either relevant or not relevant. According to Harman (1995), the pooling method is a valid sampling technique because all the participating systems would use ranked retrieval methods. The documents that are most likely to be relevant would be returned first. This pooling method actually implies sharing of relevance judgments from different participating systems.

In summary, the research reviewed above has shown the evidence of group consensus as a criterion for relevance judgment, as well as various applications that imply the group consensus concept. These ideas should be usable for relevance judgment as well, as Hill et al. (1995) point out: if the video recommendation system works, it probably would work for many other forms of information items. Given the collaborative nature of IR, Karamuftuoglu appeals for a reconsideration of "relevance" in a social informatics context: "This sort of information sharing has the important consequence of facilitating a reassessment of the concept of relevance in IR" (Karamuftuoglu, 1998, p. 1076).

This article proposes a method for assessing relevance that incorporates the group or peer consensus factor into the relevance measure. This method is called *Group Consensus Method*. The method, and its application in an experiment comparing user performance, is described in the remainder of the article.

### Group Consensus Method

The core idea of this approach is that relevance judgments should be made by: (1) people who would do real

searches, and (2) a group of people as a whole, leading to a group consensus, in addition to each individual's judgment. Users as individuals still make their personal relevance judgment. This judgment, however, will be enhanced by other people's judgments. The method is based on the belief that the meaning of texts is not only located in the mind of a private individual, but also in the space of relationships between individuals that constitute groups or communities (a similar statement has been made by Karamuftuoglu, 1998).

The term *group* in this study means its members share some common interests, which are represented by search questions. The members may not know each other. For example, if three users search for information to solve a similar problem, the three can be said to be in a group.

The Group Consensus Method consists of the following four steps: (1) pooling of all retrieved documents from all users for the same search question (queries may differ); (2) ranking the documents in the pool by the number of users who retrieved the same document using the same search question; (3) assigning a cutoff number to divide the pool into a relevant (consensus) set and a nonrelevant set; and (4) weighing each item in the consensus set by the number of users who retrieved it.

### *Pooling*

The pool contains all retrieved documents from all users for a search question. The purpose of this pooling is to obtain individual users' relevance judgments on the retrieved documents so that a group consensus can be decided. When users conduct their searches, they iterate the search process until a satisfactory result set is generated. Ideally, users would be allowed to mark or pick up only the relevant documents in the retrieved set to form the final set so that the pool would contain only the user-judged relevant ones. However, in some situations, such as in the case study reported later in this article, users would not be allowed to mark or pick up the relevant ones from the retrieved set. In such situations, an assumption was made that the user's final set would contain the best or most relevant document(s) judged by the user relating to the search question. This final set might also contain nonrelevant documents that the user does not want but cannot get rid of. Therefore, the pool may also contain nonrelevant documents judged by users. The inclusion in the pool of the user-judged nonrelevant documents should not have an effect on the Group Consensus Method because all documents in the pool are to be assessed.

This pooling method is similar to the one used in TREC (Harman, 1995): the retrieved documents from different systems/users are pooled together. However, in TREC, the relevance of a document in the pool is eventually judged by some human assessors. In this method, documents in the pool are ranked, weighted, and thus assessed by using the number of users who did searches. It does not involve any external human judges.

### *Ranking*

This is the process to evaluate the relevance of a document based on the collective judgment shared by a group. A single person's knowledge is limited, and he/she needs to learn from others. A document judged relevant by many people sharing the same question is more relevant than a single person's judgment. Therefore, each document in the pool is ranked according to how many users retrieved it. A document retrieved by most users would appear on the top of the list, and a document retrieved by least users would be at the bottom. Once this is done, duplicate documents would be removed.

This ranking method is similar to document ranking by the coordination levels, which is discussed in van Rijsbergen (1979, p. 97). The coordination level is the actual number of terms the query has in common with a document. Retrieved or matched documents can be partially ranked by the coordination level, based on the number obtained. The difference between the ranking proposed here, and the coordination level ranking is that the coordination level ranking is based on the number of query terms, while in this method the ranking is based on the number of people who retrieved the documents.

### *Cutoff Number*

The pool will be divided by a cutoff number into two parts: the consensus set (relevant set) and the nonrelevant set. The items in the consensus set are those that are retrieved by the cutoff number of users, or more. The items in the nonrelevant set are retrieved by users below the cutoff number and are, therefore, considered nonrelevant. The cutoff number can be decided by the system, depending on the number of users in the same group and the number of documents that were retrieved. The case study reported later in this paper describes how to determine the cutoff number.

### *Weighing*

The purpose of assigning a weight to a ranked document in the relevant set is to calculate its relevance score, which will be introduced later, to demonstrate how important this document is. There may be various ways to assign a weight to a ranked document in the relevant set. The simplest way, used in this study, is to use the number of users who retrieved it as a weight. For example, if a document is retrieved by 30 users on the same question, it will have a weight of 30. All nonrelevant items are assigned a weight of zero.

This Group Consensus Method determines the relevant documents and the weight (and the rank) of each document. The relevance of a retrieved document is decided by group consensus, or more specifically, by the number of users who retrieve it based on the same search question. The more users who retrieve it, the more relevant the document is considered.

For example, if 50 people performed searching tasks and the cutoff number was chosen as 25, a document would be

considered relevant only if it was retrieved by 25 or more users. Above the cutoff number, the more (up to the maximum number of the participants) people who retrieved it, the more relevant the document is considered. Otherwise, it would be dropped off as a nonrelevant item and a weight of zero would be assigned to it.

### Measure of Group Consensus Method

Correspondingly, a *relevance concurrence score* (referred to as *relevance score* hereafter) can be used to measure a user's ability to retrieve the same relevant documents that others retrieved. It is a measure of an individual's ability to retrieve those documents and only those documents that the user group as a whole retrieved.

The relevance score is a single-valued measure for performance. It is based on the total weights of relevant items an individual user obtained for a search question, the number of relevant items retrieved, and the number of nonrelevant items retrieved. It is in effect a unit weight of retrieved relevant documents.

#### Relevance Score

Given a search question  $q$ ,  $C_q$  can be defined as the collection of the documents in the consensus set for question  $q$ , and  $D_{uq}$  as the set of documents retrieved by a user  $u$  for the search question  $q$ .

The relevance score measure  $R_{uq}$  can be defined as:

$$R_{uq} = \frac{\sum_{i=1}^{|D_{uq}|} w_{uqi}}{|C_q| + (|C_q| - |D_r|) + |D_{ir}|}$$

where  $|D_{uq}|$  is the number of items in  $D_{uq}$ ;  $|C_q|$  is the number of items in  $C_q$ ;  $w_{uqi}$  is the weight assigned to the  $i$ th of all items retrieved by user  $u$  for search question  $q$ ;  $|D_r|$  is the number of relevant documents the user retrieved; and  $|D_{ir}|$  is the number of nonrelevant documents the user retrieved.

$(|C_q| - |D_r|)$  calculates the number of relevant documents the user missed (those included in the consensus set but the user did not retrieve). Both  $(|C_q| - |D_r|)$  and  $|D_{ir}|$  can be used as penalty factors. The reason to put penalty on missing some relevant documents or retrieving some nonrelevant ones is because both factors reflect the system/user's ability to control search results through queries. Even though in real situations, a user may not care about retrieving nonrelevant documents if some relevant ones are retrieved. The user also may not be concerned about those possibly relevant documents not retrieved if one or two really useful ones are retrieved.

The maximum unit weight a user could get will be:

$$\frac{\sum_{i=1}^{|D_{uq}|} w_{uqi}}{|C_q|}$$

That is, the user would have retrieved all and only relevant records ( $D_{uq} = C_q$ ). This is the ideal situation. However, more often a user would either miss some relevant ones and/or retrieve some nonrelevant ones. Those are negative factors.

$R_{uq}$  is approximately the maximum unit weight a user could obtain for the items being retrieved. The formula is quite straightforward: the score  $R_{uq}$  increases with the increase of the weights the user obtained for the search result ( $\sum w_{uqi}$ ). Given the total weights, the value of  $R_{uq}$  is related to the number of relevant and nonrelevant items the user retrieves and the number of relevant ones the user will miss. A user will get the highest  $R_{uq}$  score if  $D_{uq}$  completely equals  $C_q$ .  $R_{uq}$  will be devalued if the user retrieves nonrelevant items and/or misses some relevant ones.

For example, assuming the consensus set has 10 items ( $|C_q| = 10$ ) and the weight for these 10 item runs from 20 (retrieved by 20 people) down to 11. A user retrieved 25 items ( $|D_{uq}| = 25$ ) and five of them are relevant ones (found in the consensus set). If the weight for the five relevant items is 11, 12, 13, 14, 15, respectively, the user's relevance score can be calculated as:

$$R_{uq} = \frac{\sum_{i=1}^{|D_{uq}|} w_{uqi}}{|C_q| + (|C_q| - |D_r|) + |D_{ir}|}$$

$$= \frac{11 + 12 + 13 + 14 + 15 + 0}{10 + (10 - 5) + 20} = 1.86$$

If the user retrieved only six items and five of them are relevant ones with weights the same as the above, the score is:

$$R_{uq} = \frac{\sum_{i=1}^{|D_{uq}|} w_{uqi}}{|C_q| + (|C_q| - |D_r|) + |D_{ir}|}$$

$$= \frac{11 + 12 + 13 + 14 + 15 + 0}{10 + (10 - 5) + 1} = 4.06$$

The user gets a higher score because fewer nonrelevant items are retrieved.

The relevance score measure ( $R_{uq}$ ) defined above is a single-valued measure that uses the information obtained by the Group Consensus Method: relevance of a document, the number of relevant documents, and the weight for each relevant document. Other existing performance measures are not suitable for use in this study. These, among others, include standard and normalized recall and precision measures (Salton & McGill, 1983, Chapter 5; van Rijsbergen, 1979, Chapter 7), mean average precision (Korfhage, 1997, pp. 199–201; -Voorhees & Harman, 2000), Swets' E-measure (Salton & McGill, 1983, pp. 177–180), and Yao's (1995) normalized distance-based performance measure (*ndpm*).

The standard recall and precision are not suitable for the Group Consensus Method for two major reasons: (1) they require knowledge of all relevant documents in the system, which is not available practically; and (2) they do not make use of document weight information, which is used by the relevance score measure.

Normalized recall and precision, mean average precision, and Swets' E-measure are all based on the recall/precision concepts. They have the basic problems as with standard recall/precision. Because recall-precision are not suitable in this study, these measures are not suitable either.

Yao's (1995) normalized distance-based performance measure (*ndpm*) cannot be used because "The application of the proposed measure requires the user preference over the entire document collection that is usually not available . . ." (Yao, 1995, p. 144). The Group Consensus Method does not collect the user's preference data explicitly. Therefore, this measure cannot be used.

These measures will not be discussed in detail here because they are not the major concern of this article and the space is limited. Interested readers are referred to the related literature for details about them. However, to validate the use of the relevance score measure, scores on some of these measures, i.e., standard recall and precision, normalized recall and precision, as well as mean average precision will also be given later in Table 3, which contains the participants' performance measures in the case study.

## A Case Study

This method was applied in a user performance experiment to compare differences between different types of users. The purpose of the study was to see if the method, in place of traditional precision and recall measures, would be able to effectively reflect the performance differences among users.

It has been observed that the performance of users on IR systems varies widely (Marchionini, Dwiggins, Katz, & Lin, 1993). Several user characteristics have been found to correlate with user performance in computing tasks, such as a user's experience with a system, academic background, age, gender, and personality (Borgman, 1989; Egan, 1988). This study was designed to test the effect of four user characteristics on search performance: Educational level, academic background, native language, and computer experience. In the experiment, the participants' search performance was measured by the relevance score, as defined in this article, as well as the *time* spent to accomplish the assigned searching tasks. It was assumed that if the Group Consensus Method was effective, the differences between different types of users should be detected by the performance measures. Although users in this study were divided into different types, they were considered as one group because they conducted searches on the same search questions and these search questions were assumed to be neutral to all of them.

## Experimental Design

### Participants

A total of 56 student volunteers participated in the experiment. They included graduate students, at doctoral or master's level, undergraduate students, and high school students who were at grades level 11 to 13. Graduate students and undergraduate students were from the University of Toronto. The high school participants were from a public school in Toronto. These students represent a significant segment of users of IR systems. The distribution of the participants, in terms of the four characteristics, is described in Table 1: Distribution of Experiment Participants. The classification in the table is based on the information collected from the participants by a background questionnaire. As mentioned earlier, all participants were considered in one group in this study, when the Group Consensus Method was applied.

### Search Tasks

Participants were asked to search the DIALOG system's on-line databases on four search questions. To avoid possible biases toward a particular type of participants, search questions were mainly on the topic of business. Business is considered a field that is relatively independent of both engineering/science and humanities. The four search questions are presented in Appendix A.

### Procedure

The data collection process took about 2 to 3 hours per participant. Each participant was presented with an information sheet about the background of the research, and was briefed about the tasks that were to be performed. The participant would then be asked to fill out a user questionnaire for his or her background information and then to conduct searches on the on-line system. Before the search started, a brief written introduction to DIALOG was presented to the participant. The four search questions were also included in the Introduction. Once the participant finished reading the Introduction and felt ready to start, the

TABLE 1. Distribution of experiment participants.

Educational level	Native language		Academic discipline		Computer experience		
	English	Non-Engl.	Sci. & Engi.	Soc. & Hum.	High	Med.	Low
Graduate ( <i>n</i> = 18)	10	8	9	9	9	9	0
Undergraduate ( <i>n</i> = 14)	4	10	6	8	7	4	3
High school ( <i>n</i> = 24)	7	17	N/A*	N/A	2	9	13
Sub total	21	35	15	17	18	22	16

\* "N/A": not applicable to the high school students. They did not yet have a formal academic orientation.

participant was logged into DIALOG and began working on the search questions independently. All participants were instructed to work on the search questions one by one in order from 1 to 4. Search results were automatically logged into a text file. The starting and ending times of each search were recorded manually by the investigator.

Participants were allowed to iterate a search on a question until a satisfactory set of documents was generated. Each participant could have only one retrieved set as the answer to a search question. Participants were asked to indicate explicitly which set was the answer set. Evaluations on retrieved documents were based on all information included in a record except the full text of the article, due to a slow network connection. Not allowing full text may have had an impact on relevance judgments of retrieved records. However, because all participants were treated in the same way, it posed no bias toward a particular participant and should not have affected the test results toward a particular participant, or an experimental group.

The DIALOG *Target* command was recommended to participants. This command is considered more flexible than the *Select* command, and it can rank the search results. The command restricts the size of a hit set to a maximum of 50 items. That means, using this command, each participant's retrieved set for a search question was equal to or less than 50 documents.

#### Data Analysis

MANOVA, ANOVA, and Tukey tests were used together in this study to conduct statistical analyses on participants' search performance measures. For each user char-

acteristic (independent variable), a MANOVA was conducted first to test the existence of overall difference. In all MANOVA tests, the Wilk's Lambda criterion was used. If an overall significant difference was found, the results of ANOVA on each of the dependent variables (time and relevance score) were consulted to reveal on which measure the difference(s) was/were found. The Tukey test, one of the multiple range tests for pairwise comparisons of means, was then employed, when necessary, to compare groups regarding the same independent variable. Otherwise, the independent variable was assumed to have no effect on performance measures and no further analysis was conducted. All statistical tests were run at  $\alpha = 0.05$  level.

#### Results

##### Search Outcomes

Using the Group Consensus Method, pools for each search question were generated except on Question 1, which had only one correct item. A participant either successfully retrieved only that item or failed by retrieving nonrelevant items. For all the other three questions, Figure 1 describes the frequency distributions of the number of participants and the number of documents retrieved. The figure also includes the total number of documents in each pool and the average number of documents retrieved in the final answer set for each question by each participant.

For all the three search questions in the figure, a cutoff number was determined. To make the peer consensus more reliable, the number was set at 20, which was about 35% of 56 total participants who performed search tasks. That is, an

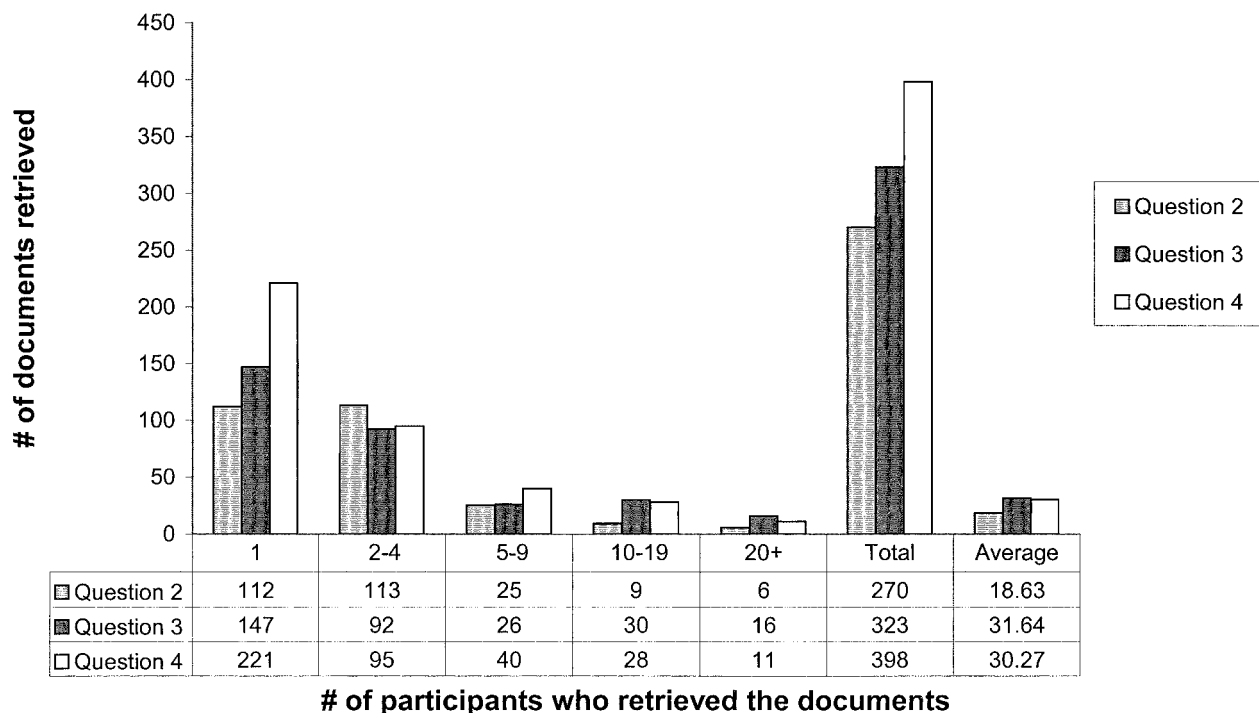


FIG. 1. Pooling of Search Results for Search Questions 2, 3 and 4.

item in a pool was considered relevant only when 20 or more participants retrieved it. Items retrieved by less than 20 participants were considered nonrelevant. For Search Question 2, there were six documents in the consensus (relevant) set. For Search Question 3, there were 16. Search Question 4 had 11. The total numbers of items retrieved by all participants (duplicates were removed) for each search question are 270, 323, and 398 for Search Questions 2, 3, and 4, respectively. The average numbers of documents retrieved by each participant for each search question are 18.63, 31.64, and 30.27. The numbers of relevant documents (those in the consensus sets) are approximately 1/3, 1/2, and 1/3 of the average answer set size for search questions 2, 3, and 4.

### Relevance Scores

The relevance score,  $R_{uq}$ , was computed, using the formula defined earlier, separately for the outcomes of each participant's searches except for Search 1. The correct answer to Search Question 1 contains only one record. The algorithm introduced earlier is not applicable to this search question. Scores for Search Question 1 were either simply "5," if the subject retrieved the correct item, or "0" if not found. Five was approximately the average score for each of the other three search questions among all participants.

Some participants could not finish all four search questions, and gave up after one or two searches. Abandoned searches could have been ignored, or not included in calculating a participant's average relevance score. However, in IR experiments, searching tasks are similar to the questions in a test for students. In a test, if a question is unanswered, the student who took the test will lose certain point(s). Similarly, in this study, abandoned searches were treated as unanswered questions on a test because abandoning a search reflected the participant's inability to accomplish that task, and this inability should be reflected in the search perfor-

mance measure: the average relevance score. Therefore, a score of zero was assigned to an abandoned search on a question. A participant might abandon a search from the beginning, without even trying, or might abandon a search after spending some time on the question but failing to retrieve relevant documents. No time was recorded for the abandoned searches in the former case because the subject did not spend time on searching. However, the time the participant spent on the failed search was recorded and included in the participant's average search time in the second case. In both cases a relevance score of zero was assigned because no relevant documents were generated.

A score for each search question was first calculated. The average time and  $R_{uq}$  scores by different types of participants for each of the four search questions are listed in Table 2. These numbers depict how the participants did on different search questions.

Although the figures in Table 2 describe the details on each search question, they are not used in the final data analysis. The relevance score  $R_{uq}$  for a participant used in the final data analysis was the average over the four search questions. In other words, it was the sum of the scores for each search question divided by four. Similar to other averages (Voorhees, 2000), this averaging may mask the differences between individual search questions, but it is necessary for the reliability of overall performance comparisons.

Mean values of performance measures from different types of users are displayed in Table 3.

In addition to the time and average  $R_{uq}$  scores, Table 3 also contains some other performance measures. These include standard recall, standard precision, normalized recall, normalized precision, and mean average precision. Formulas used for the calculation of these measures are discussed in Appendix B.

TABLE 2. Performance measures for each search question from different groups of participants.

User groups	Question 1		Question 2		Question 3		Question 4	
	Time*	$R_{uq}$ Score**	Time*	$R_{uq}$ Score**	Time*	$R_{uq}$ Score**	Time*	$R_{uq}$ Score**
Educational Level								
Graduate ( $n = 18$ )	14.11	4.44	28.72	8.63	18.11	6.07	20.49	2.65
Undergraduate ( $n = 14$ )	18.63	4.69	25.36	7.49	14.84	5.2	18.38	2.71
High school ( $n = 24$ )	31.25	4.17	40.29	5.69	17.58	4.73	11.17	2.05
First Language								
Native English ( $n = 21$ )	18.33	4.29	29.52	7.95	17.95	6.34	19.52	2.48
Non-native English ( $n = 35$ )	25.4	4.43	34.31	6.4	16.4	4.67	13.34	2.37
Academic Discipline								
Science/Engineering ( $n = 15$ )	10.72	5	30.7	11.66	19.59	5.37	19.81	2.42
Soc. sci./humanities ( $n = 17$ )	22.02	4.13	23.41	4.47	13.36	5.9	17.50	2.95
"No major" ( $n = 24$ )	31.25	4.17	40.29	5.69	17.58	4.73	11.17	2.05
Computer Experience								
High ( $n = 19$ )	13.61	5	25.94	8.4	20.44	6.59	19.56	2.67
Medium ( $n = 21$ )	24.36	4.09	31.68	7	15.55	5.36	18.36	2.52
Low ( $n = 16$ )	30.81	4.06	41.06	5.35	15.06	3.76	7.56	1.97

\* Average time (in minutes) spent on the search question by a participant in the group.

\*\* Average relevance score on the search question by a participant in the group.

TABLE 3. Mean search performance measures from different groups of participants.

User groups	Mean values		MANOVA results*	Other measures**				
	Time <sup>a</sup>	$R_{uq}$ <sup>b</sup>		$R^c$	$P^d$	$R_{norm}$ <sup>e</sup>	$P_{norm}$ <sup>f</sup>	$MAP^g$
Educational level			$F(4, 104) = 5.35, p < 0.01$					
Graduate ( $n = 18$ )	20.81	5.45		0.61	0.51	0.83	0.82	0.13
Undergraduate ( $n = 14$ )	17.43	4.93		0.60	0.49	0.82	0.80	0.12
High school ( $n = 24$ )	25.07	4.16		0.54	0.39	0.66	0.66	0.10
First language			$F(2, 53) = 0.85, p = 0.43$					
Native English ( $n = 21$ )	21.18	5.26		0.65	0.47	0.85	0.85	0.13
Non-English ( $n = 35$ )	22.38	4.47		0.53	0.45	0.69	0.68	0.11
Academic discipline			$F(4, 104) = 5.91, p < 0.01$					
Science/engineering ( $n = 15$ )	19.63	6.2		0.62	0.59	0.89	0.88	0.15
Soc. Sci./humanities ( $n = 17$ )	19.06	4.36		0.59	0.42	0.76	0.75	0.10
“No major” ( $n = 24$ )	25.07	4.16		0.54	0.39	0.66	0.66	0.10
Computer experience			$F(4, 104) = 2.08, p = 0.08$					
High ( $n = 19$ )	19.71	5.67		0.67	0.51	0.83	0.83	0.13
Medium ( $n = 21$ )	22.17	4.74		0.55	0.47	0.80	0.79	0.12
Low ( $n = 16$ )	23.61	3.79		0.52	0.37	0.60	0.59	0.09

<sup>a</sup> Average time (in minutes) spent across all 4 search questions by each participant.

<sup>b</sup> Average relevance scores across all 4 search questions by each participant.

<sup>c</sup> Standard recall

<sup>d</sup> Standard precision

<sup>e</sup> Normalized recall

<sup>f</sup> Normalized precision

<sup>g</sup> Mean average precision

\* Tests based on Time and  $R_{uq}$  scores.

\*\* Averages across Search Questions 2, 3, and 4. Calculations of these measures are described in Appendix B.

As demonstrated in Table 3, the relevance scores for different groups are consistent with all other measures that are based on the recall-precision concepts. In every user category, the ranking of the groups, from high to low in terms of their relevance scores, is the same in any other measures. Given that recall and precision are the established measures, this consistence demonstrates the meaningfulness and the validity of the relevance score measure for search performance.

However, it should be noted that the data analyses for this case study were conducted only on the average  $R_{uq}$  scores and the search time. Other measures are listed in the table only for intellectual interests and for a comparison to the relevance scores. No further data analyses were conducted on these data because the calculations of these measures have to use some estimated data, and in this case, the calculations were based on the information available from the Group Consensus Method, which might have bias in favor of the relevance score.

#### Results of MANOVA Tests

MANOVA tests found that the educational level had a significant,  $F(4, 104) = 5.35, p < 0.01$ , overall effect on search performance measures. However, no significant difference was found between the two language groups, with  $F(2, 53) = 0.85, p = 0.43$ . The results are consistent with the findings from previous user studies (e.g., Charoenkitkarn, 1996).

Academic background was found to have a significant,  $F(4, 104) = 5.91, p < 0.01$ , overall effect on search per-

formance measures. Because there were no clear discipline orientations among the high school participants (no major), there was no discipline separation of the participants in this group. Social science students used a little less average time than science/engineering students did on searching. However, science/engineering students achieved a much higher relevance score on search questions than that obtained by the social science students.

More experienced computer users generally spent less time on searching and their relevance scores were higher, as noted in Table 3. However, the MANOVA test found the differences were not significant,  $F(4, 104) = 2.08, p = 0.08$ . The result indicated that there was no significant effect by computer experience on users' search performance measures.

#### Results of ANOVA Tests

The ANOVA analyses on individual variables (time and score) found that on educational levels, the differences occurred on the variable of search time. No significant difference was found on the variable of relevance score, although there was a tendency for participants with a higher level of education to obtain higher scores.

Academic discipline had effects on both relevance score and search time.

The results are summarized in Table 4.

The relationships between science/engineering, social science/humanities and “no major” participants were tested by a Tukey test. Results of Tukey tests are presented in Table 5.

TABLE 4. Results of ANOVA for educational level and discipline.

Dependent variable	Educational level	Discipline
Time	$F = 10.65, df = 2, p = 0.0001$	$F = 8.39, df = 2, p = 0.0007$
Score	$F = 1.8, df = 2, p = 0.18$	$F = 4.8, df = 2, p = 0.01$

As noted in Table 5, the test found that in terms of time, there were differences between the graduate and high school participants and between undergraduate and high school participants. No significant difference was found between graduate and undergraduate participants. The results indicated the important role of time spent on performing searches. As Table 3 showed earlier, high school participants spent the longest time on searches. Undergraduate participants' average time was shorter than that of graduate participants. The reason might be that a few undergraduate participants did not complete all four search questions.

In terms of relevance score, science/engineering participants differed significantly from social science/humanities participants and "no major" participants. Differences were not revealed by the ANOVA analysis on educational level. The differences confirmed similar findings from some previous studies (Borgman, 1989; Kamala, 1991; Qiu, 1993). It should be noted that these studies were conducted about a decade ago. Unfortunately, no later studies have been found that are related to the findings reported here.

*Interactions of the User Characteristics on Search Performance*

A MANOVA test was performed to examine the possible effects of the interactions of the independent variables. Because undergraduate science/engineering participants were all non-native English speakers, the interaction of native language and academic background did not make much sense and was not included in the analysis. The analysis focused on the interactions between the four user characteristics. None of the interactions had a significant effect on search performance.

*Summary of the Case Study*

Educational level had an effect on participants' search performance. Specifically, differences were found in the sample data between graduate and high school subjects and between undergraduate and high school subjects. The differences were found in the amount of time the participants spent on searching. No significant difference was found between graduate and undergraduate participants.

Using about the same amount of time, science/engineering students obtained significantly higher relevance scores than social science/humanities subjects did. The finding that the two groups had different performance once again confirmed the results from some previous studies.

Participants with low-level computer experience spent longer in accomplishing searches and obtained lower relevance scores than did those with high and medium-level experience. However, the differences were not statistically significant.

There was no significant tendency for native English participants to have better search performance than non-native English-speaking participants. Nevertheless, given the importance of the language skills in IR tasks, the effect of native language should be considered further in future experiments with more participants and stricter control of background variance.

Effects of the interactions of the four user characteristics were also tested. None of the interactions between educational level, native language, academic background, and computer experience was found significant on participants' search performance.

One limitation of this experiment was the lack of random sampling when selecting participants. Instead, they were recruited as volunteers. Therefore, there might be a bias in their attitude towards using computers. Another problem related to the sample was the unbalanced allocation of subjects in native language and academic background groups, as pointed out earlier in this article. Due to these imbalances, the analyses on the interactions between independent variables were limited. More details about this case study can be found in Zhang (1998).

**Conclusions and Discussion**

This article proposes the Group Consensus Method for assessing relevance of retrieved documents. The method is based on the social informatics concepts, i.e., collaborative IR and/or interactions among users. The advantages of the method are: (1) it considers social contexts in which real searches occur; (2) it is more objective while keeping a user's subjective judgment; (3) by using it, the relevance of a retrieved document is easier to measure than traditional recall and precision measure, and (4) the relevance score can be calculated automatically, without the involvement of external human assessors.

The Group Consensus Method was applied in a user performance experiment, and the results of the experiment showed the effectiveness of the method. It appears to have

TABLE 5. Results of Tukey test for educational level and discipline.

Groups compared between:	Difference found	
	Time	Score
Graduate and undergraduate	No	No
Graduate and high school	Yes*	No
Undergraduate and high school	Yes*	No
Sci/Engi. and Soc/Hum.	No	Yes*
Sci/Engi. and no major	Yes*	Yes*
Soc/Hum. and no major	Yes*	No

\* Difference significant at  $\alpha = 0.05$  level was found between the two groups.

been adequate for the purposes of getting an objective measure. Although a significant difference was found on relevance scores in only one case, in all other cases there were apparent trends that scores increased/decreased between different types of participants. The results indicate that the method, including the relevance score measure, was quite effective in distinguishing participants' search performance. As the results showed, differences on the resulting relevance scores between types of participants were observed. The relevance score proves to be a valid measure and is able to measure users' search performance in terms of search outcome.

This method can be used in controlled IR experiments to measure user performance, as demonstrated in the case study. It also has potential implications for IR system designs. There are a number of ways in which the method may be incorporated into and thus enhance IR systems. One way is to calculate the relevance score to enhance a document's relevance ranking for a given search query. Given today's available computing power, it is possible to collect and record search information for each document and for each user in the system. The data recorded can thus be used for calculating relevance scores. The calculated relevance score information can also be provided to users for their reference. A user would benefit from such information for knowing how he or she did on the same question comparing to other users so that his/her search query could be adjusted.

However, the reliability of the method needs to be tested in more studies. The method needs to be compared with other relevance-based measures and used in more IR evaluation experiments. The case study reported in this article involved different types of users, and they were considered as one group. It will be interesting to use the method for user groups with a uniform characteristic to compare the effect of the method with that of other performance measures.

The relevance score formula used in the method is only one of many that could have been used. It is possible that some other alternative measures could be developed. Further research is required to develop improved methods for assessing relevance.

## Acknowledgments

The author thanks Professor Charles Meadow for advising on the initial idea of this method and the anonymous referees for their constructive comments. The author also gratefully acknowledges the help of Kim Ingersoll on an earlier draft of this article and the valuable suggestions made by Professor Carol Doll, Wayne State University.

## References

Barry, C.L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, 45(3), 149–159.

Barry, C., & Schamber, L. (1998). Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing & Management*, 34(2,3), 219–236.

Blair, D.C. (1996). STAIRS redux: Thoughts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science*, 47(1), 4–22.

Borgman, C.L. (1989). All users of IR systems are not created equal: An exploration into individual Differences. *Information Processing & Management*, 25(3), 237–251.

Boyce, B.R., Meadow, C.T., & Kraft, D.H. (1994). *Measurement in information science*. San Diego: Academic Press.

Charoenkitkam, N. (1996). The effect of markup-querying on search pattern and performance in large-scale text retrieval. Ph.D. Dissertation, Department of Industrial Engineering, University of Toronto.

Egan (1988). Individual differences in human-computer interaction. In *Handbook of human-computer interaction* (pp. 543–568). North Holland: Elsevier Science Publishers B.V.

Harman, D. (Ed.). (1995). *Overview of the third text retrieval conference (TREC-3)* (pp. 1–19). Gaithersburg, MD: National Institute of Standards and Technology.

Harter, S.P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9), 602–615.

Hersh, W. (1994). Relevance and retrieval evaluation: perspectives from medicine. *Journal of the American Society for Information Science*, 45(3), 201–206.

Hill, W., Stead, L., Rosenstein, M., & Furnas, G. (1995). Recommending and evaluating choices in a virtual community of use. *Proc. CHI Conference on Human Factors in Computing Systems*, Dever, CO (pp. 194–201).

Kamala, T.N. (1991). Individual differences in the use of CD ROM databases. Ph.D. dissertation, University of Hawaii at Manna, Honolulu, HI.

Karamuftuoglu, M. (1998). Collaborative IR: Toward a social informatics view of IR interaction. *Journal of the American Society for Information Science*, 49(12), 1070–1080.

Korfhage, R.R. (1997). *Information storage and retrieval*. New York: John Wiley & Sons.

Marchionini, G., Dwiggins, S., Katz, A., & Lin, X. (1993). Information seeking in full-text end-user-oriented search systems: The roles of domain and search expertise. *Library and Information Science Research*, 15, 35–69.

Meadow, C.T., Marchionini, G., & Cherry, J.M. (1994). Speculations on the measurement and use of user characteristics in IR experimentation. *Canadian Journal of Information and Library Science*, 19(4), 1–22.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9), 810–832.

Newell, S.C. (1997). User models and filtering agents for improved Internet IR. *User Modeling and User-Adapted Interaction*, 7, 223–237.

Qiu, L. (1993). Markov models of search state patterns in a hypertext information retrieval system. *Journal of the American Society for Information Science*, 44(7), 413–427.

Resnick, P., & Varian, H. (1997). Recommender systems. *Communications of the ACM*, 40(3), 56–58.

Rieh, S.Y., & Belkin, N.J. (2000). Interaction on the Web: Scholars' judgment of information quality and cognitive authority. *Proceedings of the 63rd ASIS Annual Meeting* (vol. 37, pp. 25–36).

Rioux, K. (2000). Sharing information found for others on the World Wide Web: A preliminary examination. *Proceedings of the 63rd ASIS Annual Meeting*, 37, 68–77.

Salton, G., & McGill, M.H. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26, 321–343.

Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology*, 29, 3–48.

Shardanand, U., & Maes, P. (1995). Social information filtering: Algorithms for automating "Word of Mouth." *Proc. CHI Conference on Human Factors in Computing Systems*, Dever, CO (pp. 210–218).

- Su, L.T. (1994). The relevance of recall and precision in user evaluation. *Journal of the American Society for Information Science*, 45(3), 207–217.
- Twidale, M.B., & Nichols, D.M. (1998). Designing interfaces to support collaboration in information retrieval. *Interacting with Computers*, 10(2), 177–193.
- Twidale, M.B., Nichols, D.M., & Paice, C.D. (1997). Browsing is a collaborative process. *Information Processing & Management*, 33(6), 761–783.
- Tiamiyu, M.A., & Ajiferyke, I.Y. (1988). A total relevance and document interaction effects model for the evaluation of IR processes. *Information Processing & Management*, 24(4), 391–404.
- van Rijsbergen, C.J. (1979). *Information retrieval*. London: Butterworths.
- Voorhees, E.M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5), 697–716.
- Voorhees, E.M., & Harman, D. (2000). Overview of the eighth Text REtrieval Conference (TREC-8). In *TREC-8 Proceedings*. <http://trec.nist.gov/pubs.html>.
- Yao, Y.Y. (1995). Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science*, 46(2), 133–145.
- Yee, I.H. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science*, 44(3), 161–174.
- Zhang, X. (1998). A study of the effects of user characteristics on mental models of information retrieval systems. Unpublished Ph.D Dissertation, Faculty of Information Studies, University of Toronto.

## Appendix A: Search Questions

- (1) There is an Ontario-based company that produces plastic products. The name of the company seems to be “PrecisionCraft” or something like that. Please use **File 520: “D&B-Canadian Dun’s Market Identifiers”** to find the company’s address, telephone number, number of employees, and when the company established.
- (2) Find articles about Bernardo’s psychiatric test(s) in File 262: **“Canadian Business and Current Affairs.”**
- (3) In 1989 an oil spill happened in Alaska. The oil ship *Valdez* had the worst oil spill ever witnessed in the U.S. A company named “Exxon” was charged with responsibility for the disaster. A professor whom you work for wants to know what has been reported about the Exxon case. Please use **File 148: “IAC Trade & Industry Database”** to find relevant information.
- (4) “Home banking” is called a revolution in banking industry. Among many features, telephone banking, electronic banking, and the use of smart card and debit card are the rapidly developing methods for home banking. Your client needs to review published papers that discuss the trend of home banking in the above aspects. Please use file 15: ABI/INFORM to find relevant information.

## Appendix B. Calculations of Recall, Precision, Normalized Recall, Normalized Precision, and Mean Average Precision

Calculations of standard recall and precision normally involve some estimated data. In this case study, these measures (the “Other Measures” contained in Table 3) were

calculated based on the numbers obtained through the Group Consensus Method.

Given  $q$  = a search question;  $C_q$  = the collection of the documents in the consensus set for question  $q$ ;  $D_{uq}$  = the set of documents retrieved by a user  $u$  for the search question  $q$ ;  $|D_{uq}|$  = the number of items in  $D_{uq}$ ;  $|C_q|$  = the number of items in  $C_q$ ;  $|D_r|$  the number of relevant documents the user retrieved;  $|D_{ir}|$  = the number of nonrelevant documents the user retrieved; and  $N$  = the total number of documents in the pool for a search question, the formulas for calculating the measures are as follows.

Recall:  $R$  = Number of Retrieved Relevant Items/Number of Total Relevant Items in the System =  $|D_r|/|C_q|$ .

Precision:  $P$  = Number of Retrieved Relevant Items/Number of Total Retrieved Items =  $|D_r|/|D_{uq}|$ .

Mean Average Precision:

$$\text{Average Precision} = \sum_{i=1}^{|D_r|} P_i/|C_q|.$$

For any  $|C_q| - |D_r|$  (unretrieved relevant documents),  $P = 0$ .

Normalized Recall:

$$\begin{aligned} R_{\text{norm}} &= 1 - \left( \left( \sum_{i=1}^{REL} RANKi - \sum_{i=1}^{REL} i \right) / REL(N - REL) \right) \\ &= 1 - \left( \left( \sum_{i=1}^{|D_{uq}|} w_{uqi} - \sum_{i=1}^{|D_{uq}|} i \right) / |C_q| (N - |C_q|) \right). \end{aligned}$$

where  $RANKi$  is the  $i$ th relevant item’s rank in the consensus set;  $REL$  is the number of relevant documents.

Normalized Precision:

$$\begin{aligned} P_{\text{norm}} &= 1 - \left( \sum_{i=1}^{REL} \log RANKi \right. \\ &\quad \left. - \sum_{i=1}^{REL} \log i / \log(N!/(N - REL)!REL!) \right) \\ &= 1 - \left( \sum_{i=1}^{|D_{uq}|} \log RANKi \right. \\ &\quad \left. - \sum_{i=1}^{|D_{uq}|} \log i / \log(N!/(N - |C_q|)!|C_q|!) \right). \end{aligned}$$

where  $RANKi$  is the  $i$ th relevant item’s rank in the consensus set;  $REL$  is the number of relevant documents.