JIS

Testing a decision-theoretic approach to the evaluation of information retrieval systems

Ye Diana Wang

Department of Applied Information Technology, George Mason University, USA

Guisseppi Forgionne

Information Systems Department, University of Maryland Baltimore County, USA

Abstract.

With information overload a real problem, especially on the Internet, there has been much interest in developing effective and efficient information retrieval (IR) systems. The various information retrieval approaches will require accurate evaluation to justify the requisite substantial development and implementation investment. Recently, a comprehensive and integrated evaluation model has been proposed and illustrated. By analyzing the evaluation measures using the analytic hierarchy process (AHP), the model transforms IR evaluation into a multi-criteria decision making (MCDM) problem, which assesses both the IR outcome and the interactive IR process. This paper extends that research by refining the evaluation model and by testing the research question through mathematical testing and simulation. The tests confirm the need to include both process and outcome criteria in any IR evaluation and prove the superiority of the proposed decision-theoretic approach over the traditional evaluation methodologies that focus on the IR outcome alone.

Keywords: information retrieval system; evaluation; information search process; decision making; multi-criteria model; analytic hierarchy process

1. Introduction

1.1. Problem and background

The system-centered approach to the evaluation of information retrieval (IR) systems has remained dominant for more than 40 years. However, this approach, which originated from batch-mode evaluations in controlled laboratory settings (e.g., [1, 2]), does not completely account for user–system interactions in modern IR systems during the IR process. Two major limitations make the system-centered approach incomplete from an evaluation perspective. First, the static performance measures used in

Correspondence to: Ye Diana Wang, Department of Applied Information Technology, George Mason University, Bull Run Hall, Room 120, MS 4F5, 10900 University Drive, Manassas, VA 20110, USA. Email: ywangm@gmu.edu

the evaluation, such as precision and recall, are context free and incapable of reflecting the interactive process of IR [3] and the multidimensional nature of relevance [4, 5]. This approach evaluates only the IR outcome, and ignores the IR process. The second major limitation is that the user dimension is absent from the evaluation. As a result, information needs and relevance are set by the system rather than the user [6].

In response to the limitations of the traditional system-centered approach to IR evaluations, there have been several studies which have taken alternative evaluation approaches. For instance, Su [7, 8] identified 20 measures grouped into the criteria of relevance efficiency, utility and user satisfaction, including potential underlying dimensions, and aimed to identify a single best measure. Based on these findings, Su [9, 10] proposed a model, combining traditional measures and user satisfaction measures, to be applied in realistic retrieval situations, and presented an application of the model to the evaluation of Web search engines by undergraduates. Johnson, Griffiths, and Hartley [11] defined IR evaluation as a multidimensional construct and grouped a variety of user-centered evaluation measures into four criteria of system performance. Ultimately, they provided a framework in which evaluations for each of the dimensions could potentially relate to both relevant system factors and situational impacts. More recent system evaluation incorporates real users, tasks and systems (e.g., [12]). These studies aimed to measure the end product as well as the experience of the user's information seeking process, and argued that it is impossible to evaluate the effectiveness of information access systems based on a single prototypically correct response (i.e., measure).

Although these studies differed in methodology and in the number or types of criteria or measures included, they all support the notion that IR evaluation is multi-dimensional and should be measured with reference to users in an applied context. As pointed out by Harter

> we need new approaches for measuring retrieval perfomance that do not depend on a single set of fixed, unchanging relevance assessments, and/or pooling retrieval results of many individual searches. We need to develop approaches to evaluation that are sensitive to these variations, i.e., approaches that reflected the real world of real users [13: p. 48].

Therefore, a dominant problem in current IR is the question of how to incorporate various criteria and measures within a unified and comprehensive IR evaluation based on the user's actual information retrieval experience and the user's definition of relevance.

1.2. Motivation

We began with an examination of end-users' actual information-seeking and retrieval processes. As a result of two decades of empirical research, the end-users' actual information search process (ISP) has been identified as a six-stage constructive activity–task initiation, topic selection, prefocus exploration, focus formulation, information collection, and search closure [14]. A closer investigation of these six stages indicates that the ISP is actually a decision-making process, and the activities involved in the ISP are consistent with the procedures and steps that a decision-maker normally follows in making a decision [15]. Indeed, a user's IR process is often characterized as a decision-making or problem-solving activity in a great deal of the information seeking literature (e.g., [16, 17]). According to Soergel [18], the ultimate objective of any information storage and retrieval system is to improve the task performance, problem-solving or decision-making of the user. In other words, an effective IR system can help the users to make good decisions, by effectively and efficiently locating the information relevant to their needs during their information search process. Thus, it would seem realistic that the effectiveness of an IR system should be measured in terms of its ability to facilitate the users' information search process by retrieving relevant information and its impact on the users' decision-making. Subsequently, the decision-making steps involved during users' ISP have been further identified [19].

This identification of decision-making steps offers important implications for IR evaluations. For one thing, the goal for IR evaluation is not only to measure how much the precision or recall rate goes up or down, but also to determine the effectiveness of the IR system in facilitating the users' IR interaction and improving users' decision-making by retrieving relevant information. By drawing on

decision making theories, IR evaluation can be transformed into a multi-criteria decision making (MCDM) problem, which involves the assessment of both process and outcome criteria. These motivating ideas have led to the proposal of a novel decision-theoretic approach to the evaluation of IR systems in a previous study [19]. This paper extends that research by refining the evaluation model and by proposing and testing a research question, in which the proposed decision-theoretic approach to IR evaluation is compared with the traditional system-centered approach.

The rest of the paper is organized as follows. First, the paper presents the proposed decision-theoretic approach to the evaluation of IR systems and specifies the evaluation model, with related criteria and measures, as well as the methodology for conducting the evaluation. Next, the testing of the research question, using mathematical testing and statistical simulation, is reported. The results and limitations of the study are also further discussed. The paper concludes with a discussion of research contributions and directions for future research.

2. A decision-theoretic approach

Comprehensive IR evaluation should be performed in a holistic, valid, and realistic manner. This section presents such an approach and provides explanations for the components of the decision-theoretic approach. These components include: (1) the decision-theoretic model with evaluation criteria and measures; and (2) the methodology for obtaining measurements and conducting evaluation.

2.1. The decision-theoretic evaluation model

The backbone of the decision-theoretic evaluation approach is the multi-criteria model (see Figure 1 and Table 1), which associates the various evaluation criteria and measures in a hierarchy [19]. Because the premise of the approach is that an effective IR system improves the process of a user's decision-making during information searching and, therefore, leads to better outcomes, the measures are separated into process and outcome criteria and sub-criteria in the model.

Process-oriented measures are mainly used to assess the results of human interaction with the IR system *during the search session* and are newly proposed based on the major decision-making steps identified from the ISP. Outcome-oriented measures are mainly used to assess results *after the search session* and have been created based on well-known formulae. The inclusion of both process and outcome criteria makes the proposed decision-theoretic approach different from most existing evaluation approaches, and more suitable for evaluating IR interactions [20, 21]. In a way, the model identifies the factors that must be measured to evaluate the success of an IR system from a decision-making perspective. It is also possible to isolate the specific cause of a particular decision outcome

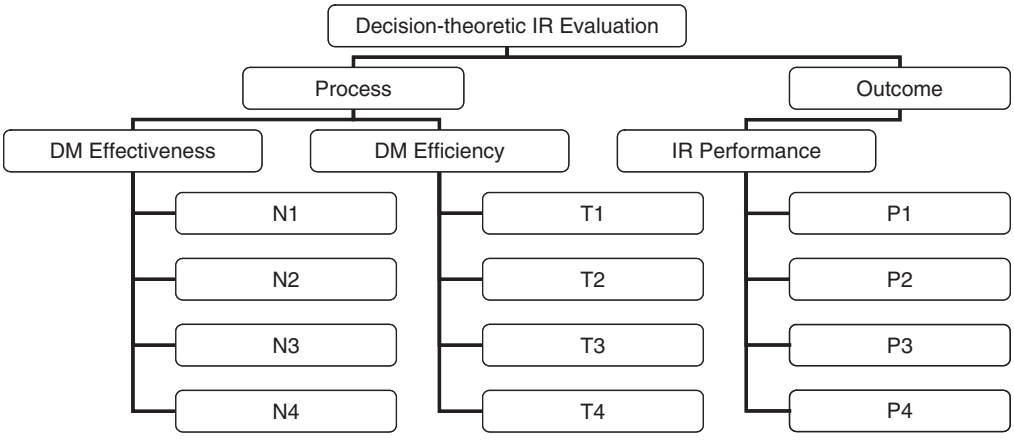

Fig. 1. Multi-criteria model for the evaluation of IR systems.

Table 1
Evaluation measures in the multi-criteria model

Measure abbreviation	Measure description
T1	Time to recognize the problem
T2	Time to identify general topic(s)
T3	Time to establish focus (search queries)
T4	Time to identify relevant documents
N1	Number of general topic alternatives
N2	Number of documents generated for the general topic(s)
N3	Number of relevant documents identified
N4	Number of additional documents identified from rechecking
P1	Utility
P2	User learning
P3	DM precision
P4	DM recall

by investigating the measures for the different decision-making steps involved in the ISP. The following sub-sections provide brief explanations for the evaluation criteria and measures included in the model. The relevant elicitation and calculation procedures (i.e., how the measurement data should be collected and measured) have been detailed and illustrated elsewhere [19].

2.1.1. Process criteria and measures

The process-oriented evaluation measures correspond to the decision-making steps previously identified within the ISP [19], and assess each step in terms of both effectiveness and efficiency. Traditionally in the IR field, efficiency and effectiveness have been seen and used as outcome-oriented evaluation measures or criteria. In a system-centered approach, efficiency is usually measured in terms of the computer resources used, such as CPU time and storage, and effectiveness is commonly measured in terms of precision and recall [22]. In a user-centered approach, it is not uncommon to use efficiency interchangeably with 'search session time' and effectiveness with 'search success' [7].

However, efficiency and effectiveness possess different meanings from a decision-making perspective and are treated as process-oriented measures in this model. *Efficiency* is related to the time needed for the user to perform each decision-making step during the IR process, such as the time to establish search queries or the time to identify relevant documents. *Effectiveness* is related to the user's decision productivity at each step during the IR process, for example the number of general topic alternatives generated or the number of relevant documents identified. For this reason, we may also name the two sub-criteria 'DM (decision-making) efficiency' and 'DM effectiveness'.

2.1.2. Outcome criteria and measures

The outcome criteria of the proposed decision-theoretic evaluation model aim to assess IR performance, that is, the overall quality of the retrieval results both to the user (utility and user learning) and the system (precision and recall). *User learning* represents the user's increase in understanding of the current problem or acquired skills for future or further decision-making as a result of the information search [23]. The *utility* measure is used to represent whatever the user finds to be of value about the system output, whether that is usefulness, appropriateness, or entertainingness [24, 25].

To adapt to the proposed decision-theoretic approach, the *precision* and *recall* measures require a reinterpretation from their traditional system-centered, batch-model evaluation definition. In the decision-theoretic approach, relevance of the retrieved documents is judged by the user who owns the information need. The user has complete control to extract and examine documents, and identify which document is relevant and useful with respect to the information need that initiated the information search. Therefore, precision and recall used in the proposed approach are revised to the forms

$$DM\ precision = \frac{number\ of\ relevant\ documents\ identified\ by\ the\ user}{number\ of\ all\ documents\ extracted\ by\ the\ user}$$

$$DM\ recall = \frac{number\ of\ relevant\ documents\ identified\ by\ the\ user}{number\ of\ all\ documents\ retrieved\ by\ the\ system}$$

where 'relevant documents identified by the user' refers to the final documents that have been judged by the user to be relevant at the end of the information search. The number of these documents is equivalent, using the process-oriented measures, to the sum of N3 (number of relevant documents identified) and N4 (number of additional documents identified from rechecking). 'All documents extracted by the user' refers to all the documents that have been extracted and examined by the user throughout the information search, some of which may be eventually identified as relevant while others may not. The number of these documents is equivalent to the sum of N2 (number of documents generated for the general topic), N3, and N4. 'All documents retrieved by the system' refers to all the documents that have been technically relevant and thus, retrieved by the system, some of which may eventually be extracted by the user while others may not. The number of such documents is often automatically returned by IR systems, especially Internet-based IR systems, as 'the total number of documents found'.

2.2. Methodology

According to Saracevic, methods or methodologies refer to

> the design, manner, means and procedures used to get and analyze evaluation results [26: p. 144]

This section presents the methodology for obtaining the measurements, including the experimental setting and data collection methods, and for analyzing the collected measurements using the analytic hierarchy process (AHP). The relevant data collection methods have been explained in detail elsewhere and illustrated using a realistic example with one user subject [19]. Hence, the focus here is on an explanation of the experimental setting in which data are collected and the application of the AHP in the decision theoretic IR evaluation.

2.2.1. Experimental setting

The experimental setting aims to facilitate IR evaluations in a way that is as close as possible to users' actual information search processes, but still in a relatively controlled test environment. The balance between realism and control in the experimental setting is achieved by two major components:

- the use of simulated work task situations, and
- the involvement of users.

The first component of the proposed experimental setting draws on the concept of a simulated work task situation, which has been proposed and tested by Borlund [27, 28]. This concept is also in line with the recently evolved trend of task-based IR evaluation [21, 29]. A simulated work task situation is a short 'cover-story' which describes a situation that may lead to information seeking. It is the 'realism and control ensuring device' in an experiment [27]. From a cognitive viewpoint, information need is seen as a user-individual and dynamic concept and originates from an *anomalous state of knowledge* or a *problematic situation* [30–33]. The simulated work task situation thus triggers a simulated information need for the user to perform information seeking tasks and develop individual need interpretation accordingly. It also specifies the context of the work domain in which the system is evaluated and serves as a platform for the user's information need development and subjective relevance assessments. All simulated work task situations are given in a stable format and require similar tasks. Therefore, experimental

#1 Simulated work task situation: Parent

Suppose that you are a concerned mother of 24-year old son who has lost three jobs in the past year. You have recently observed your son's frequent behaviors of extreme anger and are worried that these problems might be due to depression, which could lead to suicidal tendencies. You want to help your son and decide to search for relevant information on suicide prevention.

Fig. 2. Example of a simulated work task situation.

control is made possible by providing consistent treatment to all the user subjects with respect to the use of simulated work task situations, and consequently generating comparative results across both the system and the group of user subjects.

Empirical evidence shows that a simulated work task situation works most effectively when it is realistic to the test subjects; that is, when it reflects a situation with which they can identify [27, 28]. Hence, simulated work task situations should cover a variety of suitable scenarios, while being tailored to fit the group of subjects. Figure 2 shows an example of a simulated work task situation, in which a user plays a hypothetical role as a concerned mother of an adult son with suicidal tendencies and needs to search for some relevant online information on suicide prevention to help her son. This simulated work task situation was used in our previous study in evaluating a newly developed IR system in the telemedicine domain by applying the decision-theoretic evaluation approach [19]. Other examples of simulated work task situations that are applicable to the context of the evaluation are provided in Appendix A.

The second major component of the experimental setting involves user involvement throughout the IR evaluation. An important requirement of a pragmatic evaluation approach, the involvement of real users ensures the study is realistic with reference to users' interactive seeking and retrieval processes. User involvement serves a threefold purpose: (1) a user can develop an individual and subjective information need based on the simulated work task situation; (2) each user's need interpretation can evolve and mature over session time, reflecting the user-individual, dynamic nature of information need; and (3) relevance assessments of the retrieved items can be made against the information need situation by the particular user who owns the information need. Thus, users are allowed to independently control their interaction with the IR system, examine and extract relevant information, assess their progress, and determine when the search is complete. Most importantly, they can decide what is relevant and useful with respect to their own needs in an experimental setting that is naturalistic and realistic. The subjective judgments and task-based data obtained from the users are either impossible or difficult to achieve using the static system-centered IR approach.

2.2.2. The AHP analysis of the decision-theoretic IR evaluation

As the proposed decision-theoretic evaluation model (see Figure 1) demonstrates, both process and outcome criteria (with their subsequent measures) need to be assessed in the course of evaluating an IR system. Some of the evaluation measures (e.g., process-oriented measures) can be expressed as absolute numerical values, while others (e.g., utility and user learning) are subjective in nature, but can be quantified as well. Under these circumstances, it has become clear that IR evaluation, which involves the assessment of multiple criteria and sub-criteria, could be considered as a multi-criteria decision making (MCDM) problem. Therefore, the current approach employs one of the most popular MCDM methods, the analytic hierarchy process (AHP), to analyze the collected data and arrive at a numeric decision value with respect to the overall success of an IR system.

The AHP is a practical yet powerful multi-criteria decision making method that takes into account both qualitative and quantitative aspects of a decision [34]. The method is useful for its ability to provide a solid scientific method to aid in the creative, artistic formulation and analysis of the decision problem [35] and to simplify a complex decision problem by breaking it down into a series of pair-wise comparisons of alternatives with respect to a common goal [34]. Since Thomas Saaty's initial development of the AHP in the 1970s, it has had numerous applications and demonstrated

robustness across various domains, such as decision science, economics, politics, and many others [36]. For the first time, AHP is now extended to a new application area: IR evaluation in the current research.

When applying the AHP to a decision-theoretic IR evaluation, the first step of the AHP analysis, which is to break down the problem into its component parts, is accomplished by the identification of the process-oriented and outcome-oriented evaluation measures for IR evaluation. The second step of the AHP analysis, which is to structure the component parts into a hierarchy, is accomplished by the construction of the multi-criteria decision-theoretic evaluation model. In particular, the overall goal, which is placed at the top of the hierarchy, is the decision-theoretic evaluation of IR. The middle part of the hierarchy includes the criteria, sub-criteria, and decision factors (evaluation measures) for the IR evaluation. Two alternatives are compared when applying the AHP to IR evaluation: IRS1 and IRS2. IRS1 refers to searching information using the IR system under evaluation, and IRS2 refers to searching information using a different IR system that is comparable to IRS1.

The third step of the AHP analysis is to assign relative weights to the criteria and make pair-wise comparisons of the two alternatives. The pair-wise comparison judgments between the two alternatives (IRS1 and IRS2), which are based on the data collected for each measure, convert the alternatives to relative ratios for each measure. The results from the pair-wise comparisons are used to estimate the relative weight, or relative strength, of each alternative in attaining the overall goal of the hierarchy. As the final step of the AHP analysis, an overall priority for each alternative can be determined, as a numerical indication or decision value of relative IR success.

The AHP model proposed is intended to include all (user and system-based) measures of IR value. This model evaluates IR evaluation systems within a decision making context. The baseline is the current system-centered approach, and the alternative is a more integrated and comprehensive approach.

3. Research question and testing

The workability of the decision-theoretic approach has been empirically illustrated in the context of an actual user utilizing a domain-specific IR system [19]. This section focuses on the following research question: *Does the proposed decision-theoretic approach lead to a better evaluation of IR systems than the traditional system-centered approach?*

To answer this question, however, we first need to understand that the goal for IR evaluation is not to measure how much the precision or recall rate goes up or down, but to assess the success of the IR system in facilitating the users' IR interaction and improving users' decision-making by retrieving relevant information. This assessment of IR success can be quantified and expressed as a decision value, or the AHP priority, of the system under evaluation, which is determined using the AHP analysis. Thus, a higher decision value with respect to the system comparison indicates a better evaluation; that is, a more comprehensive and accurate assessment of IR systems. In other words, an evaluation approach that can capture the decision value more fully is considered superior.

Therefore, the research question can be answered by comparing the decision values produced by the decision-theoretic approach and the traditional system-centered approach with respect to the same IR system. Decision values that are equivalent will indicate that the proposed decision-theoretic approach is no better than previous approaches. The research problem can be best addressed mathematically and statistically.

3.1. Mathematical testing

3.1.1. Decision value of the decision-theoretic approach

The mathematical derivations of the decision value of the decision-theoretic approach are provided first. The process starts with pair-wise comparisons for the evaluation measure ratings (N1 – N4, T1 – T4, and P1 – P4) of the two system alternatives (IRS1 and IRS2) and converting the ratings into

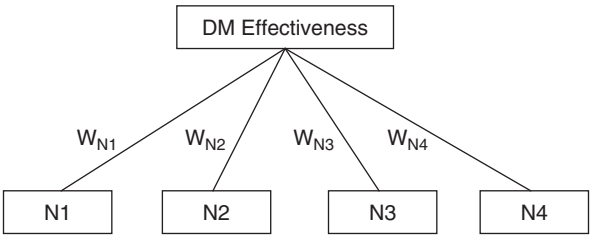

Fig. 3. The components of DM effectiveness, with weights.

relative ratios. For example, if N1 values for IRS1 and IRS2 are 20 and 4, respectively, the resulting relative ratio is $\frac{5}{6} : \frac{1}{6}$, or 0.833: 0.167. Thus, for each evaluation measure (the lowest level of the AHP hierarchy), there is a relative ratio. Because the decision value of IRS1 is the question of interest, we will concentrate on the aggregation of IRS1's decision value and use N1 – N4, T1 – T4, and P1 – P4 to represent IRS1's part of the relative ratio, which is a number between 0 and 1.

Next, to determine the potency with which the various components in one level influence the components on the next higher level, the priority of each component is calculated and aggregated from the bottom up through the hierarchy. We first illustrate the calculation of the priority of DM Effectiveness as an example. The values of N1 – N4 and their weights with respect to DM Effectiveness are as indicated along each line segment in Figure 3.

$$\textit{Priority of DM Effectiveness} = N1W_{N1} + N2W_{N2} + N3W_{N3} + N4W_{N4}$$

In a completely impartial scheme, each component would receive an equal weight at its level [15, 37]. That is,

$$W_{N1} = W_{N2} = W_{N3} = W_{N4} = W_N,$$

where W_N is the equal weight of N1 – N4 with respect to DM Effectiveness. Thus

$$\textit{Priority of DM Effectiveness} = (N1 + N2 + N3 + N4)W_N$$

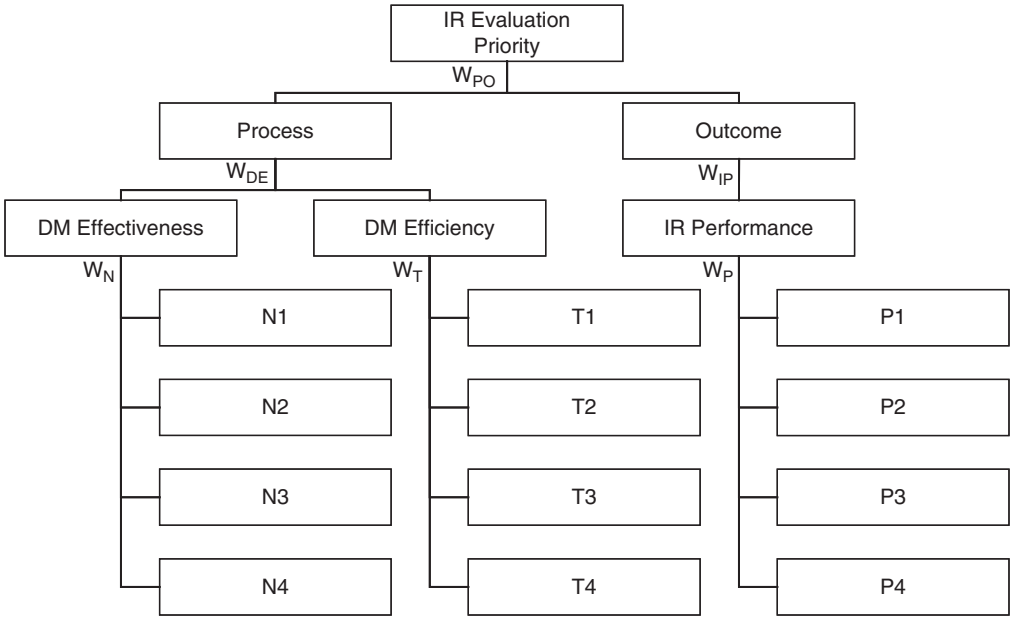

Fig. 4. The decision-theoretic approach to IR evaluation, with weights.

Next, we can move to the entire AHP hierarchy for IR evaluation (see Figure 4), where W_N, W_T, W_P, W_{DE}, W_{IP}, and W_{PO} individually represent the equal weights of the lower components with respect to the component at the higher level. We can mathematically derive the overall AHP priority by aggregating the priorities from lower level to higher level throughout the hierarchy.

The priorities of the components at the third level are:

$$\textit{Priority of DM Effectiveness} = (N1 + N2 + N3 + N4)W_N$$

$$\textit{Priority of DM Efficiency} = (T1 + T2 + T3 + T4)W_T$$

$$\textit{Priority of IR Performance} = (P1 + P2 + P3 + P4)W_P$$

The priorities of the components at the second level are:

$$\textit{Priority of Process} = (\textit{Priority of DM Effectiveness} + \textit{Priority of DM Efficiency})W_{DE}$$

$$= [(N1 + N2 + N3 + N4)W_N + (T1 + T2 + T3 + T4)W_T]W_{DE}$$

$$\textit{Priority of Outcome} = (\textit{Priority of IR Performance})W_{IP}$$

$$= [(P1 + P2 + P3 + P4)W_P]W_{IP}$$

Finally, the overall AHP priority at the first level, or the decision value of the decision-theoretic approach, is:

$$\textit{Overall AHP Priority} = (\textit{Priority of Process} + \textit{Priority of Outcome})W_{PO}$$

$$= \{[(N1 + N2 + N3 + N4)W_N + (T1 + T2 + T3 + T4)W_T]W_{DE} + [(P1 + P2 + P3 + P4)W_P]W_{IP}\}W_{PO}$$

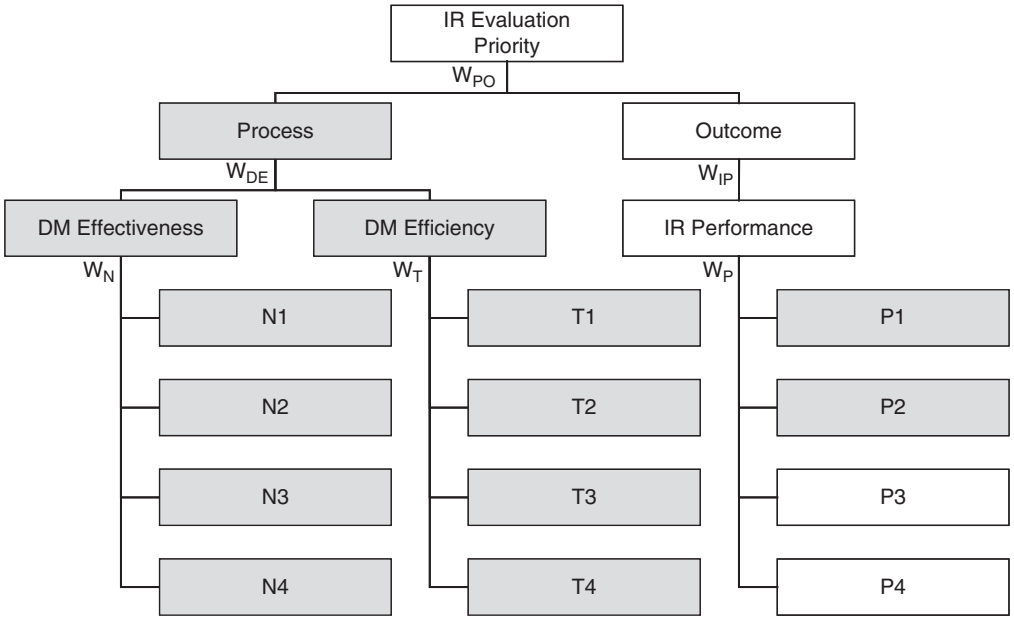

Fig. 5. The system-centered approach to IR evaluation, with weights.

3.1.2. Decision value of the system-centered approach

The mathematical derivation of the decision value of the traditional system-centered approach is relatively easy due to the small number of components involved. The system-centered approach of IR evaluation overlooks most of the assessment measures and criteria that the decision-theoretic approach is capable of capturing. The shadowed components in Figure 5 represent the overlooked evaluation measures (N1 – N4, T1 – T4, and P1 – P2), leaving only the two outcome-oriented measures used by the system-centered approach, namely *precision* (P3) and *recall* (P4).

Therefore, the priority of the component at the third level is:

$$\textit{Priority of IR Performance} = (P3 + P4)W_P$$

The priority of the component at the second level is:

$$\textit{Priority of Outcome} = (\textit{Priority of IR Performance})W_{IP}$$

$$= [(P3 + P4)W_P]W_{IP}$$

Finally, the overall AHP priority at the first level, or the decision value of the system-centered approach, is:

$$\textit{Overall AHP Priority} = (\textit{Priority of Outcome})W_{PO}$$

$$= \{[(P3 + P4)W_P]W_{IP}\}W_{PO}$$

3.1.3. Comparison of decision values

As the final step of the mathematical testing of the research question, the decision values of the decision-theoretic approach and the system-centered approach are compared by subtracting the latter from the former. Based on the current AHP hierarchy for IR evaluation, which assigns equal weights to components at the same level, the weights at each level can be calculated simply as 1/(the number of components at the associated level). Therefore, the decision value difference with definite weights is:

$$\textit{Decision Value Difference}$$

$$= \{[(N1 + N2 + N3 + N4)W_N + (T1 + T2 + T3 + T4)W_T]W_{DE} + [(P1 + P2 + P3 + P4)W_P]W_{IP}\}W_{PO} - \{[(P3 + P4)W_P]W_{IP}\}W_{PO}$$

$$= \{[(N1 + N2 + N3 + N4)W_N + (T1 + T2 + T3 + T4)W_T]W_{DE} + [(P1 + P2)W_P]W_{IP}\}W_{PO}$$

$$= \left\{\left[(N1 + N2 + N3 + N4)\frac{1}{4} + (T1 + T2 + T3 + T4)\frac{1}{4}\right]\frac{1}{2} + \left[(P1 + P2)\frac{1}{4}\right]\frac{1}{1}\right\}\frac{1}{2}$$

$$= \frac{1}{16}(N1 + N2 + N3 + N4) + \frac{1}{16}(T1 + T2 + T3 + T4) + \frac{1}{8}(P1 + P2),$$

$$\textit{Where } W_N = \frac{1}{4}, W_T = \frac{1}{4}, W_P = \frac{1}{4}, W_{DE} = \frac{1}{2}, W_{IP} = \frac{1}{1}, \textit{ and } W_{PO} = \frac{1}{2}$$

Because all the evaluation variables are numbers between 0 and 1, the resulting difference must be a positive number between 0 and 1. This result indicates that the decision-theoretic approach always provides a *higher* decision value, or *better* IR evaluation, than that of the system-centered approach.

3.2. Simulation

In practice, the proposed evaluation approach should be tested across many diverse users. Because users' ratings on the evaluation measures (N1 – N4, T1 – T4, and P1 – P4) are always expressed as

Table 2
Sample simulated data

N1	N2	N3	N4	T1	T2	T3	T4	P1	P2	P3	P4
0.362	0.021	0.866	0.708	0.204	0.060	0.999	0.325	0.026	0.894	0.960	0.186
0.943	0.589	0.211	0.537	0.235	0.775	0.073	0.073	0.382	0.340	0.482	0.696
0.171	0.576	0.959	0.128	0.900	0.255	0.650	0.551	0.383	0.232	0.227	0.299
0.810	0.481	0.966	0.685	0.924	0.432	0.100	0.003	0.733	0.429	0.877	0.602
0.299	0.018	0.766	0.551	0.169	0.980	0.586	0.379	0.753	0.216	0.032	0.060
0.725	0.403	0.522	0.874	0.250	0.128	0.677	0.405	0.731	0.156	0.585	0.282
0.164	0.459	0.369	0.691	0.608	0.622	0.742	0.199	0.243	0.884	0.424	0.180
0.502	0.723	0.410	0.466	0.879	0.169	0.691	0.819	0.969	0.386	0.431	0.242
0.844	0.530	0.521	0.204	0.637	0.462	0.576	0.882	0.018	0.941	0.761	0.485
0.868	0.752	0.750	0.524	0.239	0.070	0.110	0.382	0.216	0.033	0.008	0.435

relative ratios in a range from 0 to 1 and the actual ratings can vary considerably within the range, a simulation of many users will exhibit a wider variety of ratings than a limited sample of users. By including the entire population in the evaluation, such simulation also avoids sample representativeness problems and other confounding factors. The following sections report the simulation test of the decision-theoretic evaluation approach in comparison with the traditional system-centered approach.

3.2.1. Simulation data

The simulation was performed using the SAS statistical package [38]. The simulated population size was 10,000; that is, 10,000 simulated users were created, each with their own values of N1 – N4, T1 – T4, and P1 – P4. The variables were generated using a pseudo-random number generator formula from an assumed uniform distribution between 0 and 1. Table 2 shows the first 10 rows of simulated data, of which each row represents a user with different evaluation ratings.

For each row of simulated data, decision values were calculated, as described above, for the decision-theoretic approach and the traditional system-centered approach. Due to the large amount of data, the calculations of the decision values for the 10,000 simulated data were automated through the use of a simple computer program written in SAS. In the process, 10,000 pairs of decision values were calculated, representing the evaluation results from 10,000 simulated users using both the decision-theoretic approach and the traditional system-centered approach.

3.2.2. Simulation results

Decision value differences were tested through a two-sample paired *t*-test. In the results, DVD denotes the decision values obtained using the decision-theoretic approach, and DVS denotes the decision values obtained using the system-centered approach. The two-sample paired *t*-test is normally used to test whether the population mean of the paired differences of the two samples is significantly different from zero [39], or in our case, to determine whether the population means of DVD and DVS are equal. Equivalent population means of the decision values would indicate that *the decision-theoretic approach does not lead to a better evaluation of IR systems than the system-centered approach.*

Before performing the paired *t*-test, we must ensure that the two assumptions of the test are satisfied: the paired differences are (1) independent, and (2) identically normally distributed [39]. The first assumption is guaranteed by the way that the data were simulated. Because each row of the data, representing the ratings from an individual user, was simulated independently, the differences of DVD and DVS were also independent. The normality of the paired differences was first checked using normality tests in SAS (see Table 3). Since the sample size was greater than 2000, the Kolmogorov–Smimov (K-S) test was preferable, but the Cramer–von Mises and Anderson–Darling tests were also done, to verify the result [40]. The K-S test confirmed the normality of the sample distribution of differences ($p < 0.001$).

Table 3
Normality tests of the paired differences

Tests for normality				
Test	Statistic		*p* Value	
Kolmogorov–Smirnov	D	0.009698	Pr > D	<0.0010
Cramer–von Mises	W-Sq	0.416875	Pr > W-Sq	<0.0050
Anderson–Darling	A-Sq	2.81883	Pr > A-Sq	<0.0050

After the independence and normality of the data were confirmed, the two-sample paired *t*-test was performed. As shown in Table 4, the test revealed that there was a statistically significant difference between the population means of DVD and DVS ($t = 730.33$, $p < 0.001$).

Table 4
t-Test of the paired differences

t-test values				
Variable	Mean	Standard error	*t* Value	Pr > |*t*|
Difference	0.3754162	0.000514036	730.33	<.0001

Table 5
Summary statistics from the *t*-test

Summary statistics							
Variable	Min.	Max.	Range	Mean	Variance	Standard deviation	Standard error
DVD	0.1736	0.7975	0.6239	0.5020	0.0080	0.0892	0.0006
DVS	0.0013	0.2477	0.2464	0.1265	0.0026	0.0514	0.0004

Furthermore, the decision-theoretic approach yielded a mean decision value of 0.5020, while the mean decision value of the system-centered approach was 0.1265 (see Table 5). This finding indicates that, across the 10,000 simulated users, the system-centered approach on average captures a much smaller portion (25%, compared to 80%) of the decision value than does the decision-theoretic approach.

3.3. Discussion

The AHP puts all IR evaluation approaches on a consistent and comprehensive basis for comparison. The AHP derived decision value not only identifies the relative worth of the evaluation approaches, but also identifies the (process and outcome) sources of decision value, the benefits of the evaluation approaches, and the shortcomings of the evaluation approaches. Based on the test results, we concluded that the decision-theoretic approach gives a better evaluation of IR systems than the system-centered approach. The conclusion also confirms the need to include both process and outcome criteria in any IR evaluation. Through mathematical testing and simulation, it has been clearly shown that the system-centered approach is incomplete from an evaluation perspective, because its the ignorance of the IR process results in only a partial picture of the overall effectiveness of an IR system.

As with any research, there are limitations to this study. In the current study, these are primarily due to two assumptions. The first assumption is that each component (e.g., N1, N2, outcome, process, etc.) in the AHP hierarchy for IR evaluation should receive equal weighting at its level in terms of its relative importance. This assumption, which represents a completely impartial scheme,

reduces the number of weighting variables from 17 (each line segment from the second line down in Figure 4) to six (i.e., W_N, W_T, W_P, W_{DE}, W_{IP}, and W_{PO}), and thus significantly simplifies the mathematical calculations of the AHP priorities. We did not alter the weights and priorities because we did not want to introduce that form of bias into the analysis and no research has yet been conducted to determine the correct weightings. The second assumption is the uniform distribution of user ratings in the simulation. As the simulation is presented as an illustration that will vary across applications rather than a definitive empirical test, we assume that user ratings between 0 and 1 have an equal chance of being assigned on each evaluation measure (i.e., N1 – N4, T1 – T4, and P1 – P4). It is possible that the rates may follow different probability distributions, such as the normal or negative exponential. However, we had no prior empirical information to match with possible theoretical probability distributions. On the principle of insufficient reason (if there is no other evidence to the contrary, equal probabilities should be assigned to the events), we used the uniform probability distribution. By running sensitivity tests on additional simulations with different weight variables and probability distributions, these limitations could be overcome. The purpose of sensitivity tests is to see how the difference between the decision values of the two compared evaluation approaches changes when changes are made to the probability distributions of the simulated data. Such distributions can be generated in two ways. A pseudo-random number generation function in SAS could be used to generate a pre-assigned distribution (e.g., normal, negative exponential) containing as many points of data as are desired. Alternatively, empirical data can be collected from one or more users, and a Monte Carlo simulation method used to generate a random data set of as many observations as desired based on the empirical data. The simulated data should retain any correlations among the variables that exist in the empirical data. These additional tests are a potential area for future research.

4. Conclusions

Previously, the methods of IR evaluation proposed in the literature were fragmented and noncomprehensive. The AHP model provides a mechanism for consolidation of the fragmented measures into a unified and comprehensive model of IR evaluation. Our research aims to meet the demand for alternative IR evaluation approaches by investigating users' decision-making steps during the information search process and providing a decision-theoretic approach for the evaluation of IR systems. The authors do not refute the merits of the system-centered evaluation approach, which was developed to test IR systems in experiments involving a batch mode of processing, or its advantages in controlling all aspects of the system under evaluation. Rather, we argue for a novel evaluation approach that is based on the user's actual information retrieval experience and thus may be more suitable for assessing modern interactive IR systems. In particular, the multi-criteria model, which incorporates various evaluation measures and criteria, can be implemented through the AHP to provide a numeric decision value for the overall effectiveness of the system. In this paper, the research question, 'Does the decision-theoretic approach lead to a better evaluation of IR systems than the system-centered approach?' is answered through mathematical testing and simulation. The results confirm the need to include both process and outcome criteria in any IR evaluation and prove the superiority of the decision-theoretic approach over the traditional system-centered approach, which can capture only a partial picture of the overall effectiveness of an IR system.

Specifically, this research makes at least two important contributions to the field. First, the decision-theoretic approach transforms IR evaluation into a multi-criteria decision making (MCDM) problem, which provides a comprehensive view of both the process and outcome of IR. This approach provides a more holistic and accurate evaluation of IR systems than can be achieved with the traditional system-centered evaluation approach, which narrowly focuses on the IR outcome. Second, the research extends the analytic hierarchy process (AHP) to a new application area: IR evaluation. The AHP, which is capable of quantifying and ranking the evaluation alternatives based on simple pair-wise comparisons, offers an accessible analysis method for the IR practitioners and shows a promising future for application in the IR field. Moreover, the AHP analysis can be easily

expanded to include sub-measures of system performance. For example, the 'number of general topic alternatives' can have sub-measures such as the depth of language, the topical knowledge of the user, and/or the type of information need.

One line of future research is to investigate the alternative measures that could be used in the decision-theoretic evaluation model. As pointed out previously, the focus of the present paper is to compare the current evaluation approach with the traditional system-centered approach, and therefore it does not attempt to provide an exhaustive list of evaluation measures. Nevertheless, some alternative measures from the literature may potentially supplement the current model and overcome some of its limitations. For example, the decision-theoretic model does not differentiate between the various types of relevance (e.g., topical, cognitive, etc.) [26]. The relative relevance (RR) measure, proposed by Borlund and Ingwersen [41], describes the degree of agreement between the types of relevance applied in a non-binary assessment context, and the ranked half-RHL performance indicator denotes the degree to which relevant documents are located at the top of a ranked retrieval result. The cumulated gain (CG) and cumulated gain with discount (DCG) measures, proposed by Järvelin and Kekäläinen [42], compute the cumulative gain the users obtain by examining the retrieval results up to a given ranked position. Information seeking also has a longitudinal aspect, which may give an additional measure of IR system performance. If these additional measures were to be included in the model, they would add extra branches to the AHP hierarchy under the outcome criteria. They would also potentially lead to a different decision value, but probably not affect the overall evaluation outcome. Investigation of these alternative measures from the literature may prove fruitful in the future.

Future research may also extend the decision-theoretic evaluation approach to a wider set of IR applications in various domains, such as library or legal information retrieval. To adapt to the nature of a particular domain, new evaluation measures will need to be developed under both the process and outcome criteria after a systematic investigation of the actual IR process. In summary, future research aims at bringing new insights into the continuing development and refinement of evaluation approaches in the IR field.

Acknowledgements

The authors wish to express their gratitude to the vendors of Expert Choice for providing a free license to use the educational version of Expert Choice 11 in this research effort. Our thanks also go to the anonymous reviewers for their valuable comments, which have helped improve the paper greatly.

References

[1] C.W. Cleverdon, Aslib Cranfield Research Project: Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems (1962). Available at: https://aerade.cranfield.ac.uk/handle/1826/836 (accessed 16 March 2008).

[2] C.W. Cleverdon, J. Mills and E.M. Keen, Aslib Cranfield Research Project: Factors Determining the Performance of Indexing Systems, Vol.1: Design. Available (text and appendix) at: https://aerade.cranfield.ac.uk/handle/1826/861 and https://aerade.cranfield.ac.uk/handle/1826/862 (accessed 16 March 2008).

[3] P. Borlund, The concept of relevance in IR, *Journal of the American Society for Information Science and Technology* 54(10) (2003) 913–925.

[4] T. Saracevic, Relevance reconsidered. In P. Ingwersen and N.O. Pors (Eds), *Information Science: Integration in Perspective: Proceedings of the 2nd International Conference on Conceptions of Library and Information Science* (Elsevier, Amsterdam, 1996) 201–218.

[5] T. Saracevic, Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance, *Journal of the American Society for Information Science and Technology* 58(13) (2007) 1915–1933.

[6] G. Salton, The state of retrieval system evaluation, *Information Processing & Management* 28 (4) (1992) 441–449.

[7] L.T. Su, Evaluation measures for interactive information retrieval, *Information Processing & Management* 28(4) (1992) 503–516.
[8] L.T. Su, Value of search results as a whole as the best single measure of information retrieval performance, *Information Processing & Management* 34(5) (1998) 557–579.
[9] L.T. Su, A comprehensive and systematic model of user evaluation of web search engines: I. Theory and background, *Journal of the American Society for Information Science and Technology* 54(13) (2003) 1175–1192.
[10] L.T. Su, A comprehensive and systematic model of user evaluation of web search engines: II. An evaluation by undergraduates, *Journal of the American Society for Information Science and Technology* 54(13) (2003) 1193–1223.
[11] F.C. Johnson, J.R. Griffiths and R.J. Hartley, Task dimensions of user evaluations of information retrieval systems, *Information Research* 8(4) (2003). Available at: http://informationr.net/ir/8–4/paper157.html (accessed 15 November 2007).
[12] N. Wacholder, D. Kelly, P. Kantor, R. Rittman, Y. Sun, B. Bai, S. Small, B. Yamrom and T. Strzalkowski, A model for quantitative evaluation of an end-to-end question-answering system, *Journal of the American Society for Information Science and Technology* 58(8) (2007) 1082–1099.
[13] S.P. Harter, Variations in relevance assessments and the measurement of retrieval effectiveness, *Journal of the American Society for Information Science* 47(1) (1996) 37–49.
[14] C. Kuhlthau, *Seeking Meaning: a Process Approach to Library and Information Services* (Ablex Publishing Corporation, Norwood, NJ, 1993).
[15] G.A. Forgionne, An AHP model of DSS effectiveness, *European Journal of Information Systems 8* (1999) 95–106.
[16] L. Donohew and L. Tipton. A conceptual model of information seeking, avoiding and processing. In: Clarke, P. (ed.), *Models for Mass Communication Research* (Sage, Beverly Hills, CA, 1973) 243–269.
[17] D. Soergel, *Organizing Information: Principles of Data Base and Retrieval Systems* (Academic Press, San Diego, CA, 1985).
[18] D. Soergel, Is user satisfaction a hobgoblin? *Journal of the American Society for Information Science* 27(4) (1976) 256–259.
[19] Y.D. Wang and G.A. Forgionne, A decision-theoretic approach to the evaluation of information retrieval systems, Information *Processing & Management* 42(4) (2006) 863–874.
[20] S.E. Robertson, Process and outcome: on the evaluation of IR systems in the age of interaction, GUIs and multimedia. In: *Proceedings of Mira 99: Evaluating Interactive Information Retrieval* (1999). Available at: http://www.bcs.org/upload/pdf/ewic_mi99_paper7.pdf (accessed 12 November 2007).
[21] P. Vakkari, Task based information searching. In: Cronin, B. (ed.), *Annual Review of Information Science and Technology 37* (Information Today, Medford, NJ, 2003) 413–464.
[22] C.J. van Rijsbergen, *Information Retrieval* (2nd edn.) (Butterworths, London, 1979).
[23] G.A. Forgionne, Decision-making support system effectiveness: the process to outcome link, *Information Knowledge Systems Management 2* (2000) 169–188.
[24] W. Cooper, On selecting a measure of retrieval effectiveness, *Journal of the American Society for Information Science* 24(2) (1973) 87–100.
[25] W. Cooper, On selecting a measure of retrieval effectiveness. Part II. Implementation of the philosophy, *Journal of the American Society for Information Science* 24(6) (1973) 413–424.
[26] T. Saracevic, Evaluation of evaluation in information retrieval. In: *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM, New York, 1995) 138–146.
[27] P. Borlund, Evaluation of interactive information retrieval systems (PhD Thesis, Åbo Akademi University, Åbo, Finland, 2000).
[28] P. Borlund, The IIR evaluation model: a framework for evaluation of interactive information retrieval systems, *Information Research* 8(3) (2003). Available at: http://informationr.net/ir/8–3/paper152.html (accessed 11 May 2007).
[29] K. Järvelin and P. Ingwersen, Information seeking research needs extension toward tasks and technology, *Information Research* 10(1) (2004). Available at: http://informationr.net/ir/10–1/paper212.html (accessed 10 April 2007)
[30] N.J. Belkin, Anomalous states of knowledge as a basis for information retrieval, *Canadian Journal of Information Science* 5 (1980) 133–143.
[31] N.J. Belkin, R.N. Oddy and H.M. Brooks, ASK for information retrieval: part I, background and theory, *Journal of Documentation* 38 (1982) 61–71.

[32] P. Ingwersen, *Information Retrieval Interaction* (Taylor Graham, London, 1992).
[33] P. Ingwersen, Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory, *Journal of Documentation* 52(1) (1996) 3–50.
[34] T.L. Saaty, The analytic hierarchy process – what it is and how it is used, *Mathematical Modelling* 9(3–5) (1987) 161–176.
[35] P. Harker. The art and science of decision making: the analytic hierarchy process (Working Paper 88–06–03 Decision Sciences Department, Wharton School, University of Pennsylvania, PA, 1988).
[36] T.L. Saaty and L.G. Vargas. *Decision Making in Economic, Political, Social and Technological Environments with the Analytic Hierarchy Process* (RWS Publications, Pittsburg, PA, 1994).
[37] G.E. Phillips-Wren, E.D. Hahn and G.A. Forgionne, A multiple-criteria framework for evaluation of decision support systems, *OMEGA – The International Journal of Management Science* 32(4) (2004) 323–332.
[38] SAS Institute, *SAS System Release 8.2 (TS2M0) for Microsoft Windows* (Cary, NC, 2001).
[39] H. Motulsky, *Analyzing Data with GraphPad Prism* (1999), Available at: http://www.graphpad.com/articles/AnalyzingData.pdf (accessed 16 April 2007).
[40] G. Peng, Testing normality of data using SAS. In: *Proceedings of PharmaSUG 04* (PharmaSUG, Chapel Hill, NC, 2004).
[41] P. Borlund and P. Ingwersen, Measures of relative relevance and ranked half-life: performance indicators for interactive IR. In: B.W Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson and J. Zobel (eds), *Proceedings of the 21st ACM SIGIR Conference on Research and Development of Information Retrieval* (ACM, New York, 1998) 324–331.
[42] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In: N.J. Belkin, P. Ingwersen and M.-K. Leong (eds), *Proceedings of the 23rd ACM SIGIR Conference on Research and Development of Information Retrieval* (ACM, New York, 2000) 41–48.

Appendix A. Simulated Work Task Situations

#1 Simulated Work Task Situation: Parent

Suppose that you are a concerned mother (or father) of a 24-year old son who has lost three jobs in the past year. You have recently observed your son's frequent behaviors of extreme anger and are worried that these problems might be due to depression, which could lead to suicidal tendencies. You want to help your son and decide to search for relevant information on suicide prevention.

#2 Simulated Work Task Situation: Adult Acquaintance

Suppose that you are a co-worker of Jim, who is a 37-year old divorced man. You have recently noticed that Jim started coming in late for work and looked like he hadn't slept well. Jim started to go to happy hour everyday after work and would consistently buy rounds of drinks for everyone. You are worried that Jim's abnormal behaviors might be due to depression, which could lead to suicidal tendencies. You want to help Jim and decide to search for relevant information on suicide prevention.

#3 Simulated Work Task Situation: Employer (or Supervisor)

Suppose that you are the boss of Joe, who was one of your most consistent performers in the company for years. But over the past several months, you have noticed that Joe's productivity and focus has really fallen off. You are worried that Jim's abnormal behaviors might be due to depression, which could lead to suicidal tendencies. You want to help Jim and decide to search for relevant information on suicide prevention.

#4 Simulated Work Task Situation: Teacher

Suppose that you are a teacher of John, who is a 16-year old high-school student. In recent days, you noticed that John was disruptive in class, sullen and moody, missing assignments, and skipping school. You are worried that John's change in behavior might be due to depression, which could lead to suicidal tendencies. You want to help John and decide to search for relevant information on suicide prevention.