MEDINFO 98
B. Cesnik et al. (Eds)
Amsterdam: IOS Press

A Structured Model for Evaluation of Information Retrieval (IR)

Yuri Kagolovsky, Jochen R. Möhr

School of Health Information Science
University of Victoria, Victoria, B.C., Canada

Abstract

The paper reviews methods of evaluation of information retrieval (IR), in particular relevance (recall and precision) based approaches and their limitations and proposes a new model of IR which emphasizes the difference between the IR process and the technical IR system supporting the process. The model is proposed as basis for the structured evaluation of the IR process, and the planning, analysis and comparison of results of experiments using different methodologies.

Keywords

Information Storage and Retrieval; Evaluation Studies

Introduction

Given the emergence of Internet, digital libraries, the virtual patient record communicated through networked information systems, the general public and the health professions have been sensitized to information retrieval. In health informatics we have to develop improved approaches to retrieve information from patient records, literature, factual information systems. In order to improve methods, we have to be able to assess accomplishment. We review the state of the art of the evaluation of information retrieval and propose a solution to identified problems.

Review of the literature

The Narrow View of Information Retrieval

IR is usually understood as retrieval of any type of information from a computer [1]. Salton [2] states that IR "is concerned with the representation, storage, organization, and accessing of information items". He argues that, theoretically, there is no restriction on the type of item handled in information retrieval, a statement consistent with current Internet experience [3,4] .We concur with this view. If we use the term "document set" we use it in an inclusive manner consistent with Salton. But historically, IR research was mostly concerned with the technical aspects of the IR system. This carried through to narrow definitions of IR [5] and many researchers restrict the meaning to retrieval of information from textual databases.

Relevance Judgment Based Approaches

Consistent with this narrow definition is the use of relevance-based measures such as recall and precision in evaluation of information retrieval. Recall is the proportion of relevant documents in a retrieved document set, and precision is the proportion of retrieved document that are relevant. These measures were first proposed by Kent et al. [6] and intensively used in IR evaluation since the Cranfield experiments [7]. They are considered the "gold standard" of IR evaluation by many researchers. Although many different measures of IR system performance were proposed, relevance-based measures of recall and precision are still the most common in laboratory as well as operational settings [1,2].

However, the use of recall and precision as well as design and results of the tests were intensively criticized from the beginning of the Cranfield experiments [1,2]. This critique was mostly related to the use of relevance assessment. Salton tried to address these problems. His most important argument was that in a situation "when paired comparisons are made between various methodologies, ...the absolute performance figures of recall and precision are not of main interest. Instead, performance is judged by using the relative performance improvement of method A over method B" [3]. Thus, the argument was made that "the most solid evaluation results have been obtained with paired tests for two or more procedures carried out with otherwise fixed query and document collection" [10]. Ideally the results are evaluated by the same person.

Closer scrutiny showed that relevance judgments are subject to many influences that can affect them even during the same experiment [4] .Many experiments do not support Salton's arguments [9] .Hersh asserts that "recall and precision, may have serious problems in their external validity, at least as they are usually measured.... The controversy is not so much related to whether these concepts are important, as they obviously are, but rather to how they are used and interpreted.". Smeaton [5] analyzed the results of TREC-2 (Text REtrieval Conference) and found that even systems performing worse than average in overall performance for different searches had the best results in some specific queries. These results invalidate Salton's arguments.

Even if recall and precision could be used in paired experiments, the interpretation of their results would still be problem-

atic. One of the problems is a difference between user's information needs and their expression in queries. The second problem is that information retrieval is an interactive process and user's information needs can change during experiments. The third one is that the results of the TREC experiments disclosed the usefulness of a variation in performance characteristics of IR systems that can support different users in a variety of ways [12]. Recall and precision do therefore not sufffice to compare IR qualities.

New Models of Information Retrieval

It is obvious then, that user characteristics are important [13]. Such concepts as "relevance", "cognition", "user behavior" and "interaction" contributes largely to a "changing view of the boundaries of the system" [14]. This resulted in "a paradigmatic shift ...in the research front, to user-centered from system-centered models" [15] and to definitions of IR emphasizing cognitive, behavioral and affective aspects of IR process [16,17,18].

One of the most complete definitions [19] considers IR as "a process in which sets of records or documents are searched to find items which may help to satisfy the information need or interest of an individual or group". It defines the functions of IR, and recognizing the user and the information need as major components. The tool performing the search, the technical IR system and setting of the process are, however, not specified in this definition. A more systematic approach might therefore be desirable.

One such approach is proposed by Fidel and Soergel [2,20]. They consider the setting, user, request, database, search system, searcher, search process, and the search outcome as components of IR. For each of these elements (excluding search outcome) they provide a detailed list of variables. In another approach Tague-Sutcliffe [19] distinguishes document set, access method, user need, search strategy, the retrieved set or sequence, and the degree to which the retrieved set satisfies the user's need, the relevance judgment.

Other models vary in scope from a complete model of the functions of the technical IR system [2,20] to a model of the search of an end user for information [21]. Some of the authors [22] present the functional model of the IR process during one session, and others [23] model the information flow in the world. Existing models reflect the research interests of their authors and often consist of fragments of the IR process, representing different subsets of its structure and function. Their value for the evaluation of IR was discussed by Robertson and Hancock-Beaulieu [14]. They stress the importance of a systematic approach to IR and the defining of the boundaries of the IR system.

Consequences for Evaluation of Information Retrieval

Although different models of IR provide interesting insights and a basis for a wide variety of evaluation approaches, they do not include a synthesis of the evaluation results. Different components (elements), functions and variables of the IR process and technical IR system were identified by several researchers [2,13,14,15,19,22]. At the same time such identification does not provide a clear outline for evaluation, experimental planning and documentation. Merely knowing and controlling different variables of IR does not ensure the better understanding of the IR process. Researchers do not discriminate between "retrieval system evaluation", "retrieval system performance", evaluation of "information retrieval procedures", "retrieval evaluation" [10]. This resulted in terminological differences and incomparable results of experiments. While the existing evaluation approaches use different methodologies and terminologies, the differences are hard to identify. It would help to attain consistency in our view of the IR process. Also, rather than pursuing a quest for the perfect approach, different methods ought to be perceived as complementary.

Hersh [1] specifies two types of the evaluation of an IR system: macro-evaluation and micro-evaluation. Macro-evaluation (also viewed as clinical or field evaluations) are outcome-oriented types of evaluation as they investigate the IR system as the whole and its overall benefit. Micro-evaluations, usually performed in a controlled setting such as laboratory, serve to assess different components of the system and their impact on the performance.

A classification of Lancaster and Warner [24] defines three levels of evaluation:

1. Effectiveness of the system and the user interacting with the system. At this level the authors consider cost, time and quality, including, among other parameters, relevance-based measures of recall and precision.
2. Cost-effectiveness: This includes measures the unit costs of various aspects of the retrieval system.
3. Cost-benefit, which assesses the value of a system, the actual benefit of technology, balanced against costs of operating or using it, and addresses mainly the technical aspects.

Tague-Sutcliffe [19] raises a number of issues around the evaluation of comprehensive IR systems, including: who should make relevance judgments, real users or subject experts; how components of the process can be evaluated rather than the whole process; what kinds of aggregation are appropriate for the measures used in the evaluation of IR system; what is the value of an analytic (simulatory), as opposed to an experimental, approach in evaluating IR system; how can interactive IR systems be evaluated; what kinds of generalization are possible

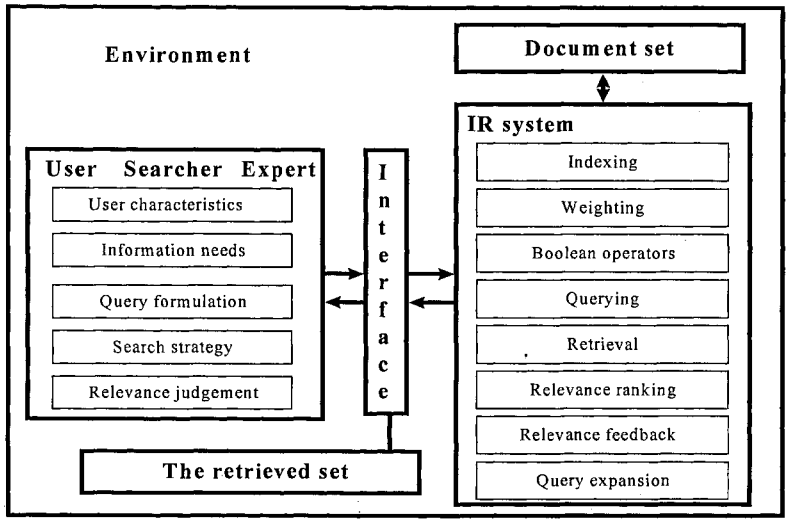

Figure 1 - Model of information retrieval (IR) process

from IR tests.

In addition, there is a growing interest in assessing the cognitive, behavioral and affective aspects through qualitative methods [17,25]. Other alternatives to evaluation of IR using relevance-based measures include investigation of the user's information needs [18,26] and assessment of user ability to find and apply specific information [27,28,29,30]. Other approaches are: the observation and monitoring of user interaction with a system [31] and "think aloud" protocol analysis [32,33,34]. There is also an example of outcome-oriented evaluation that addresses user satisfaction and system impact in health care settings [1].

A Proposal for Solution

We propose a system analytic approach distinguishing between an IR process and a technical IR system that includes identification of goals, components, structure and functions.

We define **Information Retrieval (IR)** as a process of human interaction with a technical IR system in a specific setting with the goal to find information sources relevant to a specific information need.

The IR process includes many components, functioning together to achieve some specific goals. It occurs in an environment that includes humans with specific information needs, an interface, a document set, a retrieved set, and a technical IR system. The structure and function of these components is crucial for the result (Figure 1).

All methods of evaluation from outside the boundaries in Figure 1 are macro evaluation according to Hersh [1], e.g., outcome oriented. The following discussion is concerned with micro evaluation.

The IR process is initiated by **the human being**. In some situations users will perform searches and evaluate results by themselves, in others the information needs are explained to a searcher who performs a search. In many experiments implementing Cranfield-type design and partly in TREC the prepared queries are introduced to the technical IR system in a batch mode and **the retrieved set** is evaluated by an expert. The user expresses **an information need** resulting in the processes of **query formulation** and building of a **search strategy**. All these processes depend on such **characteristics** as **state of knowledge** in specific field and **searching skills** (level of training and experience). The **environment** is characterized by such variables as kind of **setting** (laboratory or operational), **organizational policy**, or **type of research**.

The human being interacts with the technical IR system through the **interface**. The interface is any medium that transforms information queries into the system specific commands and presents the retrieved set of documents. This bi-directional function explains the relation between the processes of the query formulation by a human being and the querying mechanisms of the technical IR system. The latter work in connection with other functions of the technical IR system such as indexing of the documents in document set and queries, the weighting, Boolean operators, etc. Human beings have to be able to transform their information needs properly into the **query** (the verbalized statement of information needs) and build a search statement through the use of the interface. **The search statement** consists of the commands stated in a syntax permitted by the querying component of the technical IR system. It has a structure of search elements (terms, codes, etc.) that depends on the model of data representation in the document set built by the technical IR system.

The **technical IR** system is a computer-based information system used by human(s) during an IR process. The functions of the technical IR system include indexing, querying, weighting, Boolean operators, retrieval, relevance ranking, relevance feedback and query expansion. These interact to retrieve documents from the document set. Thus, retrieval can be generally under-

stood as the process of comparison of the search statement with the indexed documents in a document set.

The retrieved set of documents is presented to the human component of the IR process through the interface which serves as transport medium for presentation, e.g., in soft or hard copy.

Discussion

In this model a human query is entered into an IR system that performs retrieval of documents relevant to the query and returns them as a retrieved set. After this the retrieved set of documents is evaluated by a human to find documents relevant to the query or to the information needs. The need to use the evaluation of retrieved documents to evaluate the IR system changes the evaluation to one of the IR Process. Therefore, relevance-based measures of recall and precision are NOT the measures for evaluation of a technical IR system, but for evaluation of the IR process.

By proceeding according to the structure provided by the model we can make sure that the relevant components receive the necessary attention. As a computer-based information system used by a user during an IR process the technical component can be evaluated according to characteristics such as speed, data structures, etc., or as an information system with specific functions. These functions, such as indexing, querying, retrieval ranking and others can be evaluated with respect to their support of the goal of finding information sources relevant to a user's need in a database. But an IR system can function without a user, e.g., by information mapping between a CPR and a bibliographic database. In this case, the system has to be evaluated without a user's or users' relevance judgment. Therefore, an alternative evaluation of an IR system can be based on a user's queries, but will not the user's (expert's) relevance judgment for evaluation of IR results. As the TREC experiments showed there is a need to capture and analyze different patterns in the functioning of an IR system and their use to support specific user's information needs (and the purpose of a search). Examples are the variations in the number of relevant documents in the retrieved set or discovering "an obscure, tangential or 'non-obvious' semantic relationship of particular type to certain search questions" [9].

Conclusion

The structured model provides a clear specification of the components of the IR process and the technical IR system that is used in the process. It forms the basis to comprehensively evaluate the IR process in all its aspects. It allows to specify analytic and experimental methods for the investigation of IR. Existing approaches to IR evaluation can be compared and understood using this model. It also provides a basis for a structured evaluation approach to the IR process and the explicit documentation of goals, objectives and design of evaluation.

Acknowledgments

This research was supported in part by a grant from HEALNet Health Informatics and from the Province of British Columbia. We appreciate the help of Soki Kaur and Toby Walrod with preparation of this paper, and the inspiration and help with literature provided by HEALNet colleagues, in particular: David Freese, Mike Miller, Vimla Patel, and Yuri Quintana.

References

[1] Hersh WR. Information Retrieval: A Health Care Perspective. New York: Springer-Verlag New York, Inc., 1996.

[2] Salton G, and McGill MJ. Introduction to Modern Information Retrieval. New York: McGraw Hill, 1983.

[3] Kantor PB. Information retrieval techniques. Annual Review of Information Science and Technology 1994: 29 pp. 3-48.

[4] Mannoni B. Bringing museums online. Communications of the ACM 1996: 39 (6) pp. 100-5.

[5] Cleverdon C. The Cranfield tests on index language devices. Aslib Proceedings 1967: 19 pp.173-93.

[6] Kent A, Berry MM, Leuhrs FU, and Perry JW. Machine literature searching. VIII: Operational criteria for designing information retrieval systems. American Documentation 1955: 6 pp. 93-101.

[7] Cleverdon CW, and Keen EM. Factors Determining the Performance of Indexing Systems. Cranfield, UK: Aslib Cranfiled Research Project, 1966.

[8] Swanson DR. Some unexplained aspects of the Cranfield tests of indexing performance factors. Library Quarterly 1971: 41 pp. 223-8.

[9] Harter SP. Variations in relevance assessments and the measurement of retrieval effectiveness. Journal of the American Society for Information Science 1996: 47 (1) pp. 37-49.

[10] Salton G. The state of retrieval system evaluation. Information Processing & Management 1992: 28 (4) pp. 441-9.

[11] Schamber L. Relevance and information behavior. Annual Review of Information Science and Technology 1994: 29 pp. 3-48.

[12] Smeaton A. Report on TREC-2 Conference. [online] IR Digest 1993: Vol. X, No. 42, Issue 186. Available via anonymous FTP, URL: ftp://ftp.dla.ucop.edu/ in directory pub/irl/1993/

[13] Fidel R, and Soergel D. Factors affecting online bibliographic retrieval: A conceptual framework for research. Journal of the American Society for Information Science 1983: 34 (3) pp. 163-80.

[14] Robertson SE, and Hancock-Beaulieu MM. On the evaluation of IR systems. Information Processing & Management 1992: 28 (4) pp. 457-66.

[15] Tague-Sutcliffe J. The pragmatics of information retrieval experimentation, Revisited. Information Processing & Management 1992: 28 (4) pp. 467-90.

[16] Belkin NJ. Cognitive models and information retrieval. Social Science Information Studies 1984: 4 pp. 111-29.

[17] Ellis D. The dilemma of measurement in information retrieval research. Journal of the American Society for

Information Science 1996: 47 (1) pp. 23-36.

[18] Wilson TD. Information needs and uses: Fifty years of progress? In: Vickery BC, ed. Fifty years of information progress. London: Aslib, 1994; pp. 15-51.

[19] Tague-Sutcliffe J. Some perspectives on the evaluation of information retrieval systems. Journal of the American Society for Information Science 1996: 47 (1) pp. 1-3.

[20] Tague J, Salminen A, and McClellan C. A complete model for information retrieval systems. In: Bookstein A, Chiaramella Y, Salton G, and Raghavan VV, eds. SIGIR '91: Proceedings of the Fourteenth Annual International ACM/ SIGIR Conference on Research and Development in Information Retrieval (Chicago, October, 1991). New York: ACM Press, 1991; pp. 14-20.

[21] Marchionini G. Interfaces for end-user information seeking. Journal of the American Society for Information Science 1992: 43 (2) pp. 156-63.

[22] Croft WB. Knowledge-based and statistical approaches to text retrieval. IEEE Expert 1993: April pp. 8-12.

[23] Meadow CT. Text Information Retrieval Systems. San Diego: Academic Press, 1992.

[24] Lancaster FW, and Warner AJ. Information Retrieval Today. Arlington, VA: Information Resources Press, 1993.

[25] Fidel R. Qualitative methods in information retrieval research. Library and Information Science Research 1993: 15 pp. 219-47.

[26] Westbrook L. User-needs: A synthesis and analysis of current theories for the practitioner. RQ 1993: 32 pp. 541-9.

[27] Egan DE, Remde JR, Gomez LM, Landauer TK, Eberhardt J, and Lochbaum CC. Formative design-evaluation of Superbook. ACM Trans Information Systems 1989: 7 pp. 30-57.

[28] Hersh WR, Elliot DL, Hickam DH, Wolf SL, Molnar A, and Leichtenstien C. Towards new measures of information retrieval evaluation. In: Ozbolt JG, ed. Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care. Philadelphia: Hanley & Belfus, Inc., 1994; pp. 895-899.

[29] Hersh WR, Pentecost J, and Hickam DH. A task-oriented approach to information retrieval evaluation. Journal of the American Society for Information Science 1996: 47 (1) pp.50-6.

[30] Wildemuth BM, deBliek R, Friedman CP, and File DD. Medical students' personal knowledge, searching proficiency, and database use in problem solving. Journal of the American Society for Information Science 1995: 46 (8) pp. 590-607.

[31] Shneiderman B. Designing the User Interface: Strategies for Effective Human-Computer Interaction. Reading, MA: Addison-Wesley, 1992.

[32] Elstein AS, Shulman LS, and Sprafka SA. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press, 1978.

[33] Ericsson KA and Simon HA. Protocol Analysis: Verbal Reports as Data, Revised Edition. Cambridge, MA: MIT Press, 1993.

[34] Kushniruk AW, Kaufman DR, Patel VL, Levesque Y, and Lottin P. Assessment of a computerized patient record system: A cognitive approach to evaluating medical technology. M.D. Computing 1996: 13 (5) pp. 406-15.