ELECTRONIC WORKSHOPS IN COMPUTING

Series edited by Professor C.J. van Rijsbergen

Ian Ruthven (Ed)

Miro'95

Proceedings of the Final Workshop on Multimedia Information Retrieval
(Miro '95)
Glasgow, Scotland
18-20 September 1995

Paper:

Strategies for Evaluation of Interactive Multimedia Information Retrieval Systems

Nicholas J. Belkin

Published in collaboration with the
British Computer Society

Strategies for Evaluation of Interactive Multimedia Information Retrieval Systems

Nicholas J. Belkin

School of Communication, Information & Library Studies
Rutgers University
4 Huntington Street
New Brunswick, NJ 08901-1071, USA
nick@belkin.rutgers.edu

Abstract

The standard criteria for evaluation of information retrieval (IR) systems: effectiveness, efficiency, usability, satisfaction, cost-benefit seem as applicable to the interactive multimedia context as to the non-interactive, text-based context in which they have been developed. However, the operationalizations, measures and methods developed in the traditional context are, for a variety of reasons, almost wholly inadequate for the new context. This paper discusses some of the problematic aspects of evaluation in this new context, and suggests some strategies for developing new measures and methodologies for the evaluation of interactive multimedia IR systems.

1 Introduction

As we are all well aware, evaluation has been a strong suit of information retrieval (IR) and information science research from their beginnings in the 1950s. This is the case whether we consider the practice and pragmatic aspects of evaluation in IR, or if we consider the theoretical foundations of evaluation of IR systems. [1] and [2] offer excellent examples of the nature and quality of these concerns; [3] is a fine recent evaluative review of the history and nature of evaluation in IR. It is especially significant that there has been consistent attention paid to understanding and developing criteria, measures and methods for evaluation of IR systems throughout the history of IR research, and also that the IR research community has consistently insisted that high quality evaluation be a central component of any published research paper.

These characteristics of IR and information science research have led to the development of several standard criteria for evaluation, to some sets of measures which reflect these criteria, and to some standard methodologies for obtaining values on these measures. Categories of criteria which have been suggested for evaluation of IR systems include *effectiveness*, *efficiency*, *usability*, *satisfaction*, and *cost-benefit*, each with one or more different specific criteria, such as relevance or usefulness for effectiveness. For each criterion, different measures have also been suggested for their assessment (e.g. for effectiveness, recall/precision or utility), and for most of the measures, specific methodologies for their assessment have been developed (as in 'classical' IR experimentation within a test collection). Su [4] presents a thorough review of the variety of measures which have been suggested for evaluating IR system performance.

It is clear that the general categories of criteria according to which IR systems have been evaluated respond to general issues of evaluation of any system, and indeed seem to come close to exhausting the universe of such categories (a significant omission, however, is the lack of any criterion which reflects *social* acceptance or impact). In this sense, it seems that the experience to date of text-based IR system evaluation, interpreted in its broadest sense, is likely to be at least suggestive, and perhaps directly useful in the evaluation of multimedia IR systems. However, there are some aspects of how IR evaluation has been done, and there are some differences between the text-based and multimedia IR contexts, which suggest that this might not be the case.

2 Problems with the IR Evaluation Experience

Not all of the suggested sets of criteria-measures-methods triples discussed in the previous section have enjoyed equal vogue in the various communities which have been concerned with IR system evaluation. The by far dominant paradigm in IR research has been to evaluate effectiveness, in terms of retrieval of relevant documents, as measured by recall and precision, within the context of a static test collection, by comparing the performance of two or more different systems to one another, relying upon standard, constant queries and immutable, given relevance judgments. This was the model of the very earliest IR system tests in the 1950s and early 1960s, and remains the paradigmatic model to this day, as exemplified in the TREC conferences. (e.g. [5]) Although there has been a tradition of evaluation according to the other criteria, efficiency evaluation has been primarily associated with the database

community, usability with the human-computer interaction community, satisfaction with the library/information science community, and cost-benefit with IR practitioners and system designers/operators. Cross-over between communities/paradigms has unfortunately been minimal, although it is quite clear that each evaluation criterion addresses only one of the many important aspects which affect overall success of IR systems in reaching their goals. This has meant that it has been quite difficult to develop evaluation paradigms that address IR systems as wholes, as opposed to individual aspects of the systems.

A related aspect of IR system evaluation is that, by and large, it has attempted to define single measures to characterize the performance of systems. Thus, the recall/precision pair of traditional IR evaluation is often considered a problem, and much effort has been made to develop some single measure which will adequately reflect the relevance effectiveness criterion. Although satisfaction-based evaluation paradigms have usually acknowledged the necessity of multi-dimensional measurement for their criterion, (e.g. [6]) even in this tradition the quite understandable attempt has been made to find single factors which could account for results on different measures. The problem that this raises is that our experience in IR system evaluation tells us, therefore, relatively little about how to understand and take account of the significance of different facets of whatever evaluation criterion we are concerned with, even though our experience does tell us that every one that we look at is indeed multi-faceted.

Another characteristic of most evaluation paradigms for IR is that they have been developed to address success in just one kind of information-seeking behavior, that is, the retrieval of more-or-less well-specified textual information objects by searching for them in a database of a collection of such objects. It is by now commonplace to point out that this is but one form of information-seeking behavior among many, and that it is appropriate for IR systems to address these other kinds of behaviors, and to be evaluated in terms of how well they respond to them (see, e.g. [7] [8] [9]). This is clearly a significant problem if we accept this somewhat broader view of IR. For instance, at the moment we have no really good measures for evaluating the success of an IR system in its support of a user who wishes to learn about the contents of some collection of information objects by browsing through some meta-information about that collection. This is a kind of information-seeking behavior that many would now say is a legitimate concern of IR, and in any event is one which regularly occurs in a variety of circumstances.

But perhaps the most problematic aspect of IR evaluation is that most evaluation paradigms for IR have been constrained, for a variety of reasons, to addressing their problems in a non-interactive environment. Although some aspects of satisfaction-based evaluation have addressed issues of people's responses to their interactions within the IR system, and although usability-based evaluation is generally concerned with explicit measures of the ease (or difficulty) of a user's interaction within an IR system, there have been few examples of attempts to address the evaluation of the IR system as a progressive, dynamic, interactive process, in which such things as users' goals, knowledge, information problems and relevance judgments might change. It is certainly becoming recognized that this is indeed the normal IR system situation (e.g. [10]), and there have indeed been attempts to develop evaluation measures and methods which address the difficulties inherent in the interactive IR situation (see, e.g. [11] and the papers from the "interactive track" of TREC-3 in [5]). It is, however, probably fair to say that to date we have no really adequate answers to the problems raised by this situation [12].

It seems that IR evaluation has managed to finesse these problems primarily through the tactic of simplifying assumptions, a common trick for making science work. In particular, this has been done by setting highly specific goals for the IR system, from which specific criteria and measures can follow. For instance, the standard IR performance evaluation is based on a (usually) unstated goal of "satisfying an information need". This is, of course, a major simplification of the goal of IR from the user's point of view, which might more realistically be construed as "resolution of the problem which led to information seeking behavior". Clearly, the simplification of the goal leads to different possible criteria for evaluation of achievement of that goal than for achievement of the more complex one. But, even within this simplification, the criterion itself has been further simplified, to "retrieval of relevant documents", rather than, for instance, some measure of change in psychological state reflecting satisfaction of information need. Finally, the measures of effectiveness on the criterion, recall and precision, are themselves simplifications, reflecting that which can relatively easily be measured, rather than more realistic measures, such as "enough relevant documents". So the whole chain of simplifying assumptions in this case goes: the essence of "resolution of the information problem" can be captured by the "satisfaction of information need", the essence of which can be captured by "retrieval of relevant documents", whose essence can be represented by the measures of "recall" and "precision". By the time we get to the end of this chain of simplifying assumptions, we find that what we are measuring can turn out to have very little indeed to do with the highest-level user-defined goal for IR, although the argument at each step seems quite reasonable. Although this method has worked to some effect for traditional IR evaluation, it appears that the interactive multimedia IR context puts into question the validity of the entire set of simplifying assumptions. If the general technique is to work in the new context, then we at least need to re-examine the assumptions that have been made at each stage in the old context.

3 The Multimedia IR Context

For this audience, there is no real need for me to go into detail in describing the special circumstances of the context of multimedia IR, which distinguish it from text-based IR. This won't prevent me from nevertheless doing so, if only in brief, and if only as a reminder of the implications of some of these circumstances for evaluation issues.

The most obvious difference, that multimedia IR deals with information objects represented or instantiated in more than one medium, has two facets of special interest in this context. The first is that multimedia may refer to information objects which individually are in only one medium, but whose media of representation may differ from one another. This circumstance becomes important from an evaluation point of view when people's information problems can be addressed by information objects in different media, and it therefore becomes necessary to decide which (or what combination) of several representations of the equivalent information "content" (if one does not believe that the medium is the message) is the appropriate response to an information problem. In text-based IR systems, there is really no equivalent situation (which of several different levels of representation, such as title, abstract, full-text, to present seems to be the closest approximation), it being assumed that information in some textual form is always what the user desires, by virtue of having engaged in such an IR system. Although this may be a mistaken assumption, it has been the working hypothesis on which IR evaluation has been based. So, if we are for instance interested in effectiveness, it turns out that we have no experience of understanding the criteria that users might employ in evaluating a response that would be analogous to the relevance decision, but with reference to medium rather than content, nor do we have experience in evaluating the effect of different combinations of different information objects. Utility might be an overall measure which could be used in such situations, but without some way to determine and rate the components of the utility decision, this measure does not tell us very much.

The other facet of multimedia information objects that concerns us here, is that one information object may be comprised of parts in several different media. Persons might be interested, in such circumstances, in aspects of the objects which are represented in only one medium, or they might be interested in some combination of aspects, represented in different media, or they might be interested in some characterization of the information associated with the object as a whole, irrespective of medium of representation. Evaluating an IR system addressing this situation requires understanding which of these various goals the user holds, and then understanding the relationships among the different media which might affect the achievement of the goal. Furthermore, this situation suggests that evaluation according to some measure which combines the (necessarily) different measures associated with each different medium will in most cases be inappropriate; on the other hand, evaluating according to each separate measure leads to the problems of deciding which, or which combination, to choose for evaluation of any specific interaction. This seems also to be an entirely new type of problem, with which traditional IR evaluation has had no experience at all.

Although the issues discussed above are severe problems, perhaps the most significant characteristic of multimedia IR from the evaluation point of view (and in general) is its inherently *interactive* nature. In text-based IR, we have available to us a reasonably well-understood and commonly used language for describing the information objects with which one interacts; that is, *natural language*, which can serve as the meta-language for itself in either controlled or uncontrolled form. Within any specific natural language, or within any specific sub-language, there is fairly general agreement about the syntax and semantics of the language itself. Because of this general agreement, it is possible to construct fairly general representation schemes for the information objects, and it is at least possible to believe that there are fairly generally applicable ways for people to represent their information problems so that searching through comparison based on specification is a reasonable form of information-seeking behavior. Unfortunately, for many of the media in which information can be represented, this kind of agreement cannot be said to exist, except in very highly specific and constrained domains. This means that the primary form of information-seeking behavior in the multimedia environment will, for the foreseeable future, be based upon *recognition*, rather than *specification*, and on *scanning* rather than *searching*, at least when considered in fairly general environments. Information-seeking characterized by these behaviors must be accomplished interactively; we would probably characterize information-seeking of this sort as browsing, without query specification. Furthermore, because of the lack of a meta-language, and we can probably expect changes in the course of an information-seeking episode in at least the characterization of the information problem, and probably also in what is perceived as relevant. All this means that evaluation of multimedia IR will be evaluation of interactive information-seeking behavior, with all of its attendant woes. These include: having to evaluate a process, as well as (or instead of) a product; having to incorporate "real" users with "real" information problems in the evaluation design; having to evaluate with respect changeable goals and changing target specifications. Unfortunately, the IR evaluation experience does not offer much help in addressing this situation.

Thus, the multimedia IR context presents daunting problems for the evaluation of multimedia IR systems. These problems arise in multimedia IR because it is necessarily interactive, because in multimedia IR multiple goals are always possible, and multiple criteria for success are certain to exist, and because there are no obvious and general ways in which to balance, integrate or prefer one medium over another.

4. Strategies for Addressing the Problems

Given this litany of difficulties that one foresees in the evaluation of multimedia IR, and given the problems that are evident in the transfer of the text-based IR evaluation experience to the multimedia IR context, it is fair to ask if there is any hope for principled evaluation of multimedia IR at all. Obviously, the answer is a big maybe.

It appears to me that a general strategy for the evaluation of multimedia IR systems is not to try to be general. That is, many of the difficulties that arise in multimedia IR evaluation might be addressable if one limits one's concerns to the specific task and context of the specific system to be evaluated. At the moment, many, if not most multimedia IR systems are highly task-specific or context dependent. Furthermore, many of the problems that have been noted above are strongly related to the task in which the users of that specific system are engaged. These characteristics suggest that quite task-specific goals, criteria and measures need to be used for the evaluation of such systems. For instance, evaluation of effectiveness might well be measured in terms of task performance. This approach, of course, means that one gives up (at least temporarily) the goal of a general evaluation paradigm.

This approach leads to the following very general strategy for evaluation.

1. Begin with explicit task specification, which, among other things, means obtaining well-defined, user- and context-centered goals. This means explicitly that evaluation begins with studies of users in their tasks, in order to identify the criteria which they apply in evaluating success. Doing this is also a first step toward addressing such problems as balancing media.
2. Devise explicit measures for this specific system, which are directly related to the users' views of the task. This means context-specific measures. In particular, user goals in their tasks environment should lead to context-specific effectiveness and usability criteria and measures.
3. Maintain explicit separation of different media, so that each medium can be evaluated separately and so that the combinations can also be evaluated, with respect to the overall task.
4. Use cycles of evaluation methods, rather than only one method. Begin with exploratory studies to understand the parameters of the situation. Use these results to design highly controlled studies, and incorporate these results into evaluation situations in progressively less controlled environments. This will be a movement from more to less control, which should also correspond to a movement from less to more realism, in particular in terms of "real" users with "real" information problems.
5. Use different measures at each level of evaluation, rather than the same measures at all levels.
6. Triangulate, at all levels, quantitative and qualitative measures.
7. Incorporate the traditional IR evaluation experience where possible.
8. Pray that generality will emerge from experience in evaluating multiple systems.

References

1. Sparck Jones, K. (ed) *Information retrieval experiment*. Butterworths, London, 1981.
2. Harman, D. (ed) Special issue: Evaluation issues in information retrieval. *Information Processing and Management* 1992; 28: 439-528.
3. Saracevic, T. Evaluation of evaluation in information retrieval. In: Fox, E.A., Ingwersen, P. & Fidel, R. (eds) *SIGIR '95: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, 1995, pp. 138-146.
4. Su, L.T. Evaluation measures for interactive information retrieval. *Information Processing and Management* 1992; 28: 503-516.
5. Harman, D. (ed) *Proceedings of the Third Text REtrieval Conference (TREC-3)*. GPO, Washington, D.C., 1995.
6. Tagliacozzo, R. Estimating the satisfaction of information users. *Bulletin of the Medical Library Association* 1977; 65: 243-248.
7. Oddy, R.N. Information retrieval through man-machine dialogue. *Journal of Documentation* 1977; 33: 1-14.

Strategies for Evaluation of Interactive Multimedia Information Retrieval Systems

8. Bates, M. J. The design of browsing and berrypicking techniques for the online search interface. *Online Review* 1989; 13: 407-424.
9. Belkin, N.J., Marchetti, P.G., Cool, C. BRAQUE: Design of an interface to support user interaction in information retrieval. *Information Processing and Management* 1993; 29: 325-344.
10. Belkin, N.J. Interaction with texts: Information retrieval as information seeking behavior. In: Knorz, G., Krause, J., Womser-Hacker, Ch. (eds) *Information Retrieval '93: Von der Modellierung zur Anwendung*. Proceedings der 1. Tagung Information Retrieval '93. Universitätsverlag Konstanz, Konstanz, 1993, pp. 55-66.
11. Tague, J. and Schultz, R. Evaluation of the user interface in an information retrieval system: A model. *Information Processing and Management* 1989; 25: 377-389.
12. Sparck Jones, K. A look back and a look forward. In: Chiaramella, Y. (ed) *11th International Conference on Research & Development in Information Retrieval*. Presses Universitaires de Grenoble, Grenoble, 1988, pp. 13-29.