Report on Schloss-Dagstuhl Meeting on

"Evaluation of Multimedia Information Retrieval"

14-18 April 1997

edited by

Alan F. Smeaton
Dublin City University

Introduction

Information retrieval (IR), and IR tasks like information filtering and categorisation, have a long tradition of implementation and empirical evaluation based on attempting to computationally replicate users' relevance judgements. This usually involves implementing indexing and retrieval tasks on a test collection of documents and running a test collection of queries with known relevance judgements against this data. Information retrieval functionality has been evaluated quantitatively by computing how well a system performs against these known relevant documents and measuring this performance in terms of precision and recall. This classical approach to evaluation assumes a test design that resembles retrieval in batch mode. Performing experiments in artificial and synthetic environments, away from real users and real user interactions, has been the dominant mode of operation for empirical information retrieval research for decades.

Recently, information retrieval has expanded its scope to include, among others, indexing and retrieval of non-text (multimedia) documents based on external descriptors (captions) or on properties of the raw data itself. In addition, documents themselves are no longer homogeneous either in size, structure, organisation or style. As computing power increases, more computationally intensive techniques have been applied based on natural language processing, neural networks, machine learning, user modelling, intelligent and adaptive interfaces, new HCI paradigms, etc. All this means that evaluating the performance and effectiveness of an information retrieval application in terms of its precision and recall as done traditionally and in benchmarking exercises such as TREC (the DARPA-sponsored Text Retrieval Conference in which many of our participants take part), is becoming dated and unsuitable given the changed nature of the information itself. Additionally, retrieval is now highly interactive, and with multimedia documents, interactiveness plays an even more important role.

There is a clear *gap* in the field if information retrieval is to remain healthy and able to cope with the change in the nature of information and the nature of the user-system interaction. Yet the most important criterion for IR theories and IR research will always be the retrieval effectiveness of implementations, which can be validated only empirically and unless evaluation of modern information retrieval can handle this there is a fundamental deficiency in the field.

The purpose of the Dagstuhl-seminar on "Evaluation of Multimedia Information Retrieval" was to address the unsuitability of using the traditional evaluation metrics by bringing together researchers from different but overlapping and very related fields; empirical IR, HCI and user modelling. It is only by bringing together such people, in one place and at one time for an extended workshop, than we hope to make the breakthrough needed with consensus so that we can start to evaluate multimedia information retrieval as it should be done.

The format of our meeting was to divide each session into a "lead presentation" followed by a set of shorter, focussed presentations. This was then followed by either a set of group meetings of 5-6 people in each group, or a plenary discussion. We also had a number of "boiling pot" sessions where the theme for discussion emerged during the week. This loose format of organising the schedule worked well for a Dagstuhl-seminar like ours where we tend to ask more questions than provide real answers. The abstracts of the presentations made during the week are included in this booklet. The real impact of the Dagstuhl will be felt in time.

- Norbert Fuhr, University of Dortmund
- Alan F. Smeaton, Dublin City University
- Keith van Rijsbergen, University of Glasgow

Using Expected Search Length to Evaluate Multimedia IR
Mark Dunlop, University of Glasgow

In my talk I will present a new model of predictive evaluation for information retrieval. The Expected Search Duration method is based on two functions. The first function estimates, given a number of relevant documents a user wishes to find, how many documents (s)he must view to find that number. This function is based closely on Cooper's Expected Search Length. The second function takes a number of documents the user has to view and predicts how long it will take him/her to work through the given number of documents. This prediction is based on a model of the system's interface using Green and Benyon's Entity Relationship Diagrams for Interface Artefacts.

Towards the Automatic Construction of Multimedia Documents
Lynda Hardman, CWI Amsterdam

We present an approach for generating hypermedia presentations from multimedia information items distributed around a network. Our goal is to create a media-independent description of a presentation, from which multiple final presentations can be generated, taking into account the user's information need, the user's task and network and end-user platform resources.

In order to generate the structure of a hypermedia presentation from existing media items we need to define a way of grouping similar items and making links among the groups. This grouping can be based on semantic annotations attached to the media items. Current approaches to video annotation, as a complex example, are analysed. A number of research questions arising from our approach are discussed.

The ESCHER Database Editor
Lutz Wegner, Univ. Kassel, Germany

This presentation will introduce ESCHER, a database editor which supports visualization in non-standard applications in engineering, science, tourism and the entertainment industry. ESCHER was originally based on the extended nested relational data model and is currently extended to include object-relational properties like inheritance, object types, integrity constraints and methods. It serves as a research platform into areas such as multimedia and visual information systems, QBE-like queries, computer-supported concurrent work (CSCW) and novel storage techniques.

In its role as a Visual Information System, a database editor must support browsing and navigation. ESCHER provides this access to data by means of so called fingers. They generalize the cursor paradigm in graphical and text editors. On the graphical display, a finger is reflected by a colored area which corresponds to the object a finger is currently pointing at.

In a table more than one finger may point to objects, one of which is the active finger and is used for navigating through the table. The talk will mostly concentrate on giving examples for this type of navigation and will discuss some of the architectural needs for fast object traversal and display.

ESCHER is available as public domain software from our ftp site in Kassel. The portable C source can be easily compiled for any machine running UNIX and OSF/Motif, in particular our working environments IBM RS/6000 and Intel-based LINUX systems. A porting to Tcl/Tk is under way.

Showing Users Precision and Recall Figures

Gene Golovchinsky, FXPAL

The richness of the traditional information retrieval measures such as recall and precision fail to capture the interaction that pervades hypermedia and multimedia interfaces. The problem is two-fold: the measures are based on a binary model of relevance, and they confound software system performance with user behavior. Both problems stem in part from the inherent interactivity of multimedia. People react differently to different media, making context-free assessment of relevance a challenge.

While most users will derive some benefit from an image or an audio track associated with a news story, for example, the degree of the effect is quite user-dependent. Thus measures derived from assessments of relevance that do not take into consideration users' preferences for media types may not reflect accurately the actual usefulness of the retrieved documents.

Recall and precision measures, when applied to the complete man-machine system, fail to distinguish between users' behavior and software algorithm performance. Recall and precision measures were originally designed to assess the effectiveness of retrieval algorithms when processing complex queries against large databases. Interactive use, on the other hand, tends to be iterative rather than batch in nature. Users may start with a vague understanding of the search topic and then tend to refine it during the search session until the required information is found. Thus it is not clear which set of documents (those accumulated over the entire session? the ones retrieved by the last query? by the best query?) should be used to measure recall and precision.

In addition, users frequently do not examine search results exhaustively. Thus although potentially relevant documents may have been retrieved by a particular query, if they were not ranked high enough, the user may never see them. How search results are presented to users may also impact their ability to examine the retrieved documents. It is useful, therefore, when assessing man-machine system performance, to separate measures of the retrieval capabilities of the underlying search engine from measures of users' behavior. For example, while traditional recall and precision measures are used to characterize the search engine, the effectiveness of the interface in supporting the user may be measured by calculating recall and precision over the set of documents viewed as a result of the search.

This viewed subset of the retrieved set of documents can give rise to viewed recall and viewed precision measures. These measures capture in part the effectiveness of the search engine, and in part the effectiveness of the interface in exposing users to the results of the search. In the presence of explicit or implicit relevance judgments, judged recall and precision measures may also be defined. Thus the set of articles for which a user has made relevance judgments may be used to compute recall and precision measures. These measures reflect subjects' propensity to make judgments and their effectiveness in identifying relevant documents.

These measures were used to evaluate users' behavior in two experiments with dynamic hypertext interfaces. Correlation between measures was low enough to suggest that they do capture different aspects of the interaction. Thus these measures may provide additional insight into users' behavior when interacting with hypermedia information retrieval systems, increasing our ability to evaluate their use.

Interface Design Issues For Interactivity In IR
Micheline Beaulieu, City University, London

The talk will consider the relationship between three interdependent HCI factors: functional visibility, cognitive load and balance of control and explore how these can impact on search interaction in information retrieval. The discussion will draw on interface design issues raised by a set of experiments on interfaces to support query expansion in the Okapi retrieval system. Some questions to be raised include: How to integrate best elements of browsing and querying within the search task? How to support both procedural and conceptual aspects of the interactive process? What would constitute an appropriate model for interactive searching?

Personal Information Agents: Design And Evaluation Issues
Giorgio Brajnik

My talk focussed on evaluation issues that are relevant to the design of personal information agents, i.e. programs capable of supporting and amplifying user's capabilities to autonomously search an information database and get the required information. I concentrated on two scenarios: in the first one the agent is applied to a centralized bibliographic database, while in the second one the agent searches the World Wide Web (a distributed, full-text, dynamic database). Both scenarios have been explored through prototype systems that reached different development stages.

In the former scenario, the FIRE project hosted an evaluation experiment [sigir97] that demonstrated that users of information retrieval systems have to face problems that range from usability of the interface to strategic issues in carrying out a successful search session. We are now developing a new program on the basis of results derived from that experiment that is based on the following features:

- flexible sharing of activities between user and program: the user mantains the control of the session and can interactively modify the request while the program can possibly compute (in background) temporary results (like performing a 'zoom', or browsing automatically a portion of a thesaurus);
- coaching capability, where the program is able to detect a predefined set of critical situations for the user (e.g. being stuck in a difficult-to-escape query, repeated inconsistent changes to a query) and react to them. Reaction can be based on providing suggestions (like suggesting a new search strategy -- a journal run, for example) or by providing additional information (like terms deriving from a zoom operation, or from following occurrence-based 'related-term' relationships).

The design of such a system poses a number of challenging issues with respect to evaluation. Besides performance indexes also user satisfaction need to be acquired and analysed; certain design choices can be evaluated in isolation, whereas others have to depend on the entire set of features. Furthermore, user satisfaction obviously depends also on how critical situations are overcome; yet previous experiments showed that strategic problems are not perceived at all by users, who feel overwhelmed by problems of recalling and selecting appropriate terms. This aspect makes evaluation of strategic features of the system very difficult to formulate and carry out.

In the second scenario, the IfWeb system is being developed to assist a user in dynamically searching the web. IfWeb, based on the information filtering paradigm, assumes that a user profile can be automatically constructed and that it is relatively persistent over time. The system learns such a profile, uses it to select HTML pages that contain, or can lead to, relevant information items. Evaluation of such a system, needed to compare effectiveness of different features (like methods used to construct the profile, to update it as the user's interest changes, to select pages to be followed, etc.) is made difficult by a couple of issues that are related to the information resource being accessed. There is no 'retrieval' in the strict sense of the term, since the database is not centralized and there are no indexes. Therefore the system has to locate where a relevant information can be. In addition, relevance/utility of selected pages may depend only on data acquired during current or past searches, making the use of traditional term weighting schemas uneffective.

Framework for Work Analysis and System Design
Annelise Mark Pejtersen, Risoe, Denmark

It is a general feature of many evaluation experiments that it is difficult explicitly to define what functional features have actually been tested, and how comprehensive the test has been. The main purpose of this talk is to discuss how a framework for work analysis and system design can be used to structure different evaluation approaches by offering a set of compatible boundaries for planning experiments in the laboratory as well as at the user's work place. The following topics will be addressed:
•A distinction between two major approaches to evaluation •The analytical approach and the empirical approach. •Complementary analytical and empirical methods such as usability testing and cognitive walk through. •The use of the framework for empirical evaluation is illustrated by discussions of experimental setup and examples of actual evaluation experiments andtechniques. •The need to consider evaluation as a complex and dynamic process of design and evaluation.

Analysis of Real-Life requests for images
Raya Fidel, University of Washington, Seattle

Evidence suggests that the evaluation of sets of retrieved images be guided by the task for which the images are retrieved. A previous study showed that the distribution of attributes used by subjects to describe images changed with the task subjects were asked to perform. An analysis of 100 real-life requests for images showed yet another distribution. These results suggest that tasks that require image retrieval are on a spectrum with two poles: (a) the textual pole, where images are source of information; and (b) the image pole, where images are objects. Various attributes related to retrieval and evaluation receive opposite values on

opposite poles, e.g., the effectiveness of relevance feedback, ease of browsing a set. Therefore, different evaluation criteria should be used for different tasks.

Improving our Research Synergy
Marion Crehange, LORIA, Nancy, France

This presentation is centred on attempting to improve the IR research synergy by contributing to the definition of *common models* and discussing an *abstraction level* in m,odelling. We propose to use this schema recursively for objects (documents, surrogates, queries, thesauri, …) and for actions (task, relevance judgement, evaluation, …). In our approach, abstraction may contain meaning, media, etc while context may represent factors such as user, media, viewpoint, domain knowledge, interaction or task. As two persons talking together adjust their abstraction levels, here D and Q may adjust their own abstraction levels. In turn this metaphor may be considered a "concept" for a new abstraction schema, in order to make explicit the different facets of this very multi-dimensional problem.

The reverse of this, synthesis from different abstractions towards concepts, will also have to be followed. In particular, from partial evaluations one may try to manifest the construction of more global ones. The drive for these ideas is to give a means to making explicit as many factors as possible in IR evaluation and so to situate each study, each experimentation and each model with respect to others, and to common models, and possibly to improve their synergy.

Interactive Knowledge Discovery
Josiane Mothe, IRIT, Toulouse, France

The amount of information available throughout the Internet or through specific collections is so huge that more and more sophisticated information handling systems are necessary to exploit it. In addition to efficient retrieval engines, the users need some tools to be able to analyse the relevant information without having to read all of it.

Knowledge Discovery Systems main objective is to turn some selected pieces of raw information into knowledge or generalized patterns. Such a process includes a lot of problems to solve all over the three knowledge discovery phases: information harvesting and selection, information mining, results displaying. I could present the interactive method we propose to achieve knowledge discovery. It is based on information harvesting, homogenization and filtering. The discovery process itself is achieved making several modules cooperate : different mining functions and visualization modules dynamically interact -directed by the user. I could also present the operational software we developed through some screen copies or a demonstration.

Dynamic Retrieval Strategies For Interactive Image Retrieval
Adrian Müeller IBM, Germany

We will sketch an application of the logic-based multimedia retrieval system MIRACLE to image retrieval. In contrast to typical fixed-strategie approaches, MIRACLE/image is based on

- low-level feature extraction and indexing methods
- statistical evaluation of the selectivity of the low-level methods
- individual, dynamic aggregation of these basic indices and
- conceptual feedback by means of ad-hoc evaluation per current (sub)task.

Results can be maintained to provide profiles for certain groups, individuals and tasks

Design of Laboratory Experiments
Stephen Robertson, City University, London

I propose to give an update on the design of the TREC interactive track, and use that to make some observations on experimental design. One of the characteristics of TREC is that (as well as being a laboratory experiment) it's distributed; this has particular implications for experimental design. By the time of the Schloss-Dagstuhl meeting we *may* have been able to do some of the planned validation experiments (to validate the method to be used in the substantive interactive track experiment).

Evaluation of the Epic Photograph Retrieval System
Joemon Jose, Robert Gordon University, Aberdeen

We will describe a photograph retrieval system called Epic, developed at the Robert Gordon University. In Epic we use spatial features to characterise photographs. Currently, we are in the process of designing an evaluation strategy. We have built a collection of 800 photographs for conducting retrieval experiments. In this talk, I will discuss our approach to evaluation and the problems we have encountered in the selection of tasks, subjects and measures. As we are hoping to complete the evaluation before the Dagstuhl workshop, we will be able to present preliminary results at the workshop.

Applying Knowledge Of Journalist's Practices On Textual And Pictorial IR
Kalervo Jarvelin & Eero Sormunen, University of Tampere, Finland

The results of a study concentrating on needs, behaviours and practices of newspaper journalists in writing articles and searching image databases are discussed. The work reported is based on preliminary results of two projects, one developing picture retrieval methods based on automatically extracted visual attributes and manually attached textual descriptions of image contents and the other developing text retrieval based on marking journalistic content categories (e.g., agent, object, action) in the newspaper articles.

Image Retrieval via Statistical properties
James Allan, University of Massachusetts, Amherst

We have been interested in retrieving images using statistical properties of the image rather than actually trying to "understand" the elements of the picture. We have applied these ideas in color-based retrieval, in a different way though to similar effect as many other places. We have also applied it to appearance-based retrieval and achieved quite successful results, including a good face-recognition system which "knows" nothing about faces. Another intriguing applications is the retrieval of handwritten documents using techniques based upon this idea. In our approach, an OCR-like approach is used to zone the handwritten document into words which are then treated like images.

An Empirical Approach to User Interface Design
Christa Womser-Hacker, University of Regensburg, Germany

In my presentation, I report about the WING-IIR project. Although the information items consist for the most part of facts and series of measurement, the concepts of multimedia and multimodality are of interest at the user interface level.

The focus of the WING-IIR, Werkstoffinformationssystem mit Natürlichsprachlicher/Graphischer Benutzungsoberfläche und Intelligentes Information Retrieval (Materials Information System with Natural Language/Graphical User Interface and Intelligent Information Retrieval) project run by the Information Science Department of the University of Regensburg (funded by the German Ministry for Economics) lies in the empirical foundation of natural human-computer-interaction (HCI) and the practical and theoretical problems arising from the simultaneous application of mixed modalities. The fact that human communication itself consists of a mixture of different modes (e.g. natural language, deixis etc.) suggests a combination of various natural communication modes in HCI as well.

The methodological approach of rapid prototyping based on large-scale empirical investigations requires application domains operating under real world conditions. The databases and knowledge involved were made available by various cooperation partners. Different prototypes of database access methods were implemented and tested aiming at a multimodal system which combines as many of the advantages of the different access modes as possible and at the same time avoids the disadvantages. On the basis of empirical results, this initial multimodal prototype was redesigned with respect to recent software ergonomical principles, such as context sensitivity and object orientation. This new prototype is an interactive system in which the user has different possibilities of formulating his query: to fill in forms, to enter natural language queries, to „draw" the query via manipulating graphs. The kernel system is expanded by intelligent retrieval components.

Retrieval from Heterogeneous Information Sources
Tore Bratvold, UBILAB, Zürich

Information needs are often best satisfied by information from more than one database or retrieval system. However, users should not have to access one by one a large number of information sources, and then filter and merge the results manually. We aim towards a retrieval environment that provides a single access point for retrieval across multiple heterogeneous information sources and across information of different structure and media. There are hard problems to be addressed in querying (formulation, translation and distribution), in merging and ranking of results (different formats and scoring scales), and in user interaction.

We will discuss principal problems in this area, present preliminary design ideas, and draw comparison with other projects with related aims.

Measures For Interactive Multi-Modal IR and Should we Retrieve Lists of Objects ?

Nick Belkin, Rutgers University, New Jersey

There are two aspects to this presentation:

- Measures for the evaluation of interactive, multi-modal IR. The problem that needs to be addressed here is that traditional IR measures are clearly inadequate for the interactive situation, and we have no terribly good ones to replace them. I will probably suggest that any measures in such circumstances should be task-based, rather than general, as is being tried in the interactive track of TREC. I will also follow up on the experience of evaluating a visualization tool, as reported in SIGIR 96.

- Should the results of retrieving multi-media objects on the basis of different criteria (e.g. shape, text, meaning, color) be integrated into single lists/displays, and if so, how; or, should they be presented separately, and if so, how? The issue here is that we don't know much about the criteria against which successful retrieval will be judged, nor do we know much about how to represent and integrate multi-media objects for the purpose of supporting judgement processes of the user.

The Role Of Genre In Information Retrieval

Hinrich Schuetze, Xerox PARC

The notions of "text genre" and "document genre" have a crucial role to play in the development of most applications that deal with texts at a high level, including systems and schemes for the management of large-scale heterogeneous document and text repositories. We offer a brief characterization of these notions, outline their importance to various kinds of applications, and discuss the possibilities for automating generic classification.

How much Explanation do Retrieval Results Need? The Conversational Relevance of Multimedia Objects

Ulrich Thiel, GMD Darmstadt

Whereas text retrieval can - to some extent - be based on topical relevance, the problem of retrieving relevant multimedia objects is even more complex. This is due to two differences: The gap between the syntacical features and semantical interpretations is widened, and non-textual documents, i.e. pictures, videos etc. mean different things to different users.

A system which is intended to go beyond content-based retrieval methods has to incorporate a model of relevance which takes these aspects into account. Our approach employs a notion of

relevance that addresses the need for explaining why an item was retrieved, and embeds it into a framework of conversational interaction.

Evaluation Of A Multimodal Interface For Information Retrieval
Adelheit Stein, GMD Darmstadt

This explorative study primarily investigates the interaction support provided in the SPEAK! prototype. SPEAK! is a conversational interface to an information retrieval system (INQUERY, accessing a large database in the domain of art history), which is able to generate context-dependent spoken help and meta-dialogues for guiding users through the retrieval interaction. In order to evaluate the use and effectiveness of the system-generated help comments in the larger context of the interaction, 12 novice and expert users of retrieval systems were studied in their use the prototype. Two versions of the system, featuring spoken vs. written output of the generated comments, were tested. Each subject (with no preliminary training) was to solve four quite complex retrieval tasks with one version and an additional task with the other version while thinking aloud and commenting on the interaction and problems encountered. To explore the users' subjective interpretations of the interaction we mainly relied on observation and interview techniques and on qualitative analyses of the paper protocols taken and of post-experimental interviews.

Results from the observations and interviews indicate that both groups performed quite effectively and that the subjects perceived the help information produced as useful and relevant to most situations. The subjective assessments of other aspects of the system such as retrieval functionality and graphical interaction also did not vary significantly between the two groups. However, novice and expert users of retrieval systems judged differently in some respects, e.g., the novice users tended to be less critical than expert users (in particular, concerning the retrieval functionality) and, asked for a direct comparison of the two system versions, most novice users favored the spoken output mode.

Information Retrieval Evaluation: Lifebelt, Security Blanket or Straitjacket?
David J Harper, Robert Gordon University, Aberdeen

Since the earliest days of information retrieval research, we in the IR community, unlike our colleagues in many sub-disciplines of computing science, have been concerned with experimental evaluation of our ideas. In this paper, we consider IR evaluation from three viewpoints. First, we examine how evaluation has contributed to the development of our discipline. We review traditional IR approaches, and the assumptions underlying them, and we summarise the considerable achievements thereof. Second, we take a more sceptical point of view, and we highlight the acknowledged limitations of IR evaluation. We question whether our insistence on evaluating IR ideas is necessarily a good thing; perhaps this is stifling innovation and originality. So, finally, we argue that traditional IR evaluation, which includes the ideas about test collections and performance measures, may be preventing paradigm shifts in IR. Specifically, I will argue using simple examples of "new" ideas, that traditional IR evaluation may have put us in an intellectual straitjacket. By freeing ourselves from this, at least for a time, we may create a better climate in which new ideas might flourish.

Functional Description of Querying
Thomas Roelleke, University of Dortmund

Expressing an information need requires the formulation of a query. A user interface supports this formulation and the formal query is evaluated by applying the FUNCTIONAL CAPABILITIES of the underlying retrieval engine. What are the functional capabilities?

Classical IR interfaces allow the specification of a set of terms, perhaps providing logical connectors, perhaps allowing the addressing of specific fields, perhaps supporting relevance feedback. Typical DB interfaces allow the specification of attribute values and predicates for comparing the query with the retrieved objects. We want to discuss an abstraction level for describing the functional capabilities of a query interface. The functions of querying are many-folded: first, we want to describe content; also, we want to refer to specific parts of the documents and we want to use vague predicates for comparing attribute values; furtheron, we want to refer to the classification of documents and their logical structure; we want to exploit relevance feedback and we want to describe the representation of the retrieval result. The aim is to achieve a formal definition of functional capabilities for querying which can be mapped onto user interfaces and onto the underlying retrieval engine. The expressiveness of different user interfaces and different retrieval engines becomes comparable.

ERROR: undefined
OFFENDING COMMAND: their

STACK: