SYSTEMS ANALYTIC APPROACH TO THE EVALUATION OF INFORMATION RETRIEVAL (IR)

by

Yuri Kagolovsky, MD
Dnepropetrovsk Medical Institute, Ukraine, 1983

A Thesis Submitted in Partial Fulfilment of the

Requirements for the Degree of

MASTER OF SCIENCE

In the School of Health Information Science

We accept this thesis as conforming to the required standard

[redacted]

Dr. J. Moehr, Supervisor (School of Health Information Science)

[redacted]

Dr. F. Lau, Outside Member (Faculty of Business, University of Alberta)

[redacted]

Dr. H. Muller, Outside Member (Department of Computer Science)

[redacted]

Dr. A. Kushniruk, External Examiner (Department of Mathematics and Statistics, York University)

University of Victoria

ABSTRACT

Supervisor: Dr. Jochen Moehr

There is currently a high demand for an improvement to the information retrieval (IR) of documents from a variety of electronic resources. This improvement is focused on increasing the level of user satisfaction. While historically, the improvement of IR has been mainly concerned with improving the performance of search engines, new directions in research lean toward the development of a more user-centered evaluation of IR. At the same time, the variety of available evaluation methodologies, in addition to terminological differences, present difficulties for planning, analysis, and results comparison in IR research.

This thesis examines information retrieval and its evaluation methodologies using the systems approach. Problems are identified based on an overview of IR, the terminology of the field, and IR evaluation. Several different approaches to IR evaluation, including the traditional Cranfield paradigm, as well as new user-oriented paradigms, are presented. A central concept of information science, relevance, and evaluation measures using this concept, are discussed.

The systems approach to IR and its evaluation permits the identification of components of IR, their boundaries, structure, functions, and interactions. A

new definition of information retrieval and different models of IR are proposed. The systems approach and the models of IR presented here create a basis for introducing a common terminology, as well as new approaches to IR evaluation. A critique of relevance based measures of recall and precision is presented, as well as an alternative evaluation methodology that uses methods of cognitive psychology for semantics capturing and comparison. Differences between the proposed and currently used evaluation methods are discussed, and directions of future research are identified.

Examiners:

[redacted]

Dr. J. Moehr, Supervisor (School of Health Information Science)

[redacted]

Dr. F. Lau, Outside Member (Faculty of Business, University of Alberta)

[redacted]

Dr. H. Muller, Outside Member (Department of Computer Science)

[redacted]

Dr. A. Kushniruk, External Examiner (Department of Mathematics and Statistics, York University)

TABLE OF CONTENTS

ABSTRACT ..... **II**

TABLE OF CONTENTS ..... **IV**

LIST OF FIGURES AND TABLES ..... **VI**

ACKNOWLEDGEMENTS ..... **VII**

DEDICATION ..... **VIII**

CHAPTER 1: INTRODUCTION **1**

MOTIVATION 1
RESEARCH ISSUES 3
LIMITATIONS 4
RESEARCH GOALS 5
APPROACH AND METHODS 5
SUMMARY OF THE SUBSEQUENT CHAPTERS 8

CHAPTER 2: AN OVERVIEW OF INFORMATION RETRIEVAL **10**

TERMINOLOGY 10
Defining "Information Retrieval" *11*
What is an IR system? *17*
a) Definitions 17
b) Boundaries 19
Summary *21*
HISTORICAL OVERVIEW OF INFORMATION RETRIEVAL 23

CHAPTER 3: AN OVERVIEW OF IR EVALUATION **28**

AN OVERVIEW OF EVALUATION IN GENERAL 28
What is evaluation? *29*
The reasons for evaluation *30*
Goals of evaluation *31*
Types of IR evaluation *31*
Problems in IR evaluation *34*
CURRENT IR EVALUATION APPROACHES 38
Traditional IR evaluation: The Cranfield model *38*
a) Cranfield 1 1
b) Cranfield 2 41
c) Summary of the Cranfield studies 43
d) The importance of the Cranfield experiments 44
e) Large scale experiments using the Cranfield model 44
(1) The STAIRS experiment 45
(2) The TREC experiments 45
f) Problems with the Cranfield model 48
Relevance-based measures of recall and precision *49*
New Paradigms of Information Retrieval Evaluation *55*
A CONCEPT OF "RELEVANCE" 59
Importance of the concept of "relevance" *59*
Reviews *60*
Different views *61*
Problems *65*
SUMMARY 68

CHAPTER 4: THE SYSTEMS ANALYTIC APPROACH **70**

MODELS OF INFORMATION RETRIEVAL 72
General Characteristics of a Modelling Process *72*
A Search Engine Model *73*
A Formal Model of a Search Engine *75*
An Information Retrieval Process Model *76*
A Model of an Information Flow in the World *77*
A Model of the User *78*
BOUNDARIES 79
VARIABLES 81
EVALUATION FRAMEWORKS 84
CONCLUSION 86

CHAPTER 5: CONCEPTUAL IR FRAMEWORK **88**

IDENTIFICATION OF IR COMPONENTS, THEIR BOUNDARIES, AND RELATIONSHIPS 88
PROPOSED MODELS OF IR 92

CHAPTER 6: IMPLICATIONS **95**

A DEFINITION OF IR 95
AN INTRODUCTION OF COMMON TERMINOLOGY 95
RELEVANCE-BASED MEASURES OF RECALL AND PRECISION 97
An introduction *97*
Two views of a search engine *98*
The theoretical basis of recall and precision measures *99*
A critique *99*
CRITICAL ANALYSIS OF THE CONCEPT OF "RELEVANCE" 102

CHAPTER 7: PROPOSAL FOR SOLUTION **106**

A PROPOSAL FOR EVALUATING "RELEVANCE" RELATIONSHIPS 108
A PROPOSAL FOR IMPROVED EVALUATION OF SEARCH ENGINES 110
Evaluating search engines and not the whole IR process *113*
COMPARISON OF SEARCH ENGINES 119
POSSIBLE APPROACHES TO THE CAPTURING AND COMPARISON OF SEMANTICS 120
DIFFERENCE BETWEEN THE PROPOSED AND CURRENTLY USED EVALUATION METHODS 123

CHAPTER 8: CONCLUSIONS **125**

SUMMARY OF THE THESIS 125
CONTRIBUTIONS 125
FUTURE RESEARCH 126

REFERENCES **129**

APPENDIX A: GLOSSARY OF SYSTEMS CONCEPTS **146**

VITA **148**

PARTIAL COPYRIGHT LICENSE **150**

List of Figures and Tables

Figure 1. Components of the IR process 93
Figure 2. Process model of IR 94

Table 1. Possible approaches to the external evaluation of search engines 112

ACKNOWLEDGEMENTS

I would like to acknowledge the following people who helped me in the process of completing this thesis:

Dr. Jochen Moehr for being my mentor since I came to the University of Victoria, for the many things he has taught me, and for his support and encouragement during all these years. Drs. Francis Lau, Hausi Muller, and Paul Fisher for their recommendations on improving the quality of my work and help in finishing this thesis on time. Dr. Andre Kushniruk for taking the time and effort to be the external examiner. HEALNet (Health Evidence Application and Linkage Network) for providing financial support and unlimited opportunities to learn. Toby Walrod, Soki Kaur, and David Freese for their support, friendship, and help whenever I needed it. The late Mike Miller for an exchange of ideas and critique of my early work. Karen Solie for helping to improve my writing skills and for her editing services. Susan Gimbel for giving me energy to fight my way through problems. Leslie Wood and Carolyne Swayze for their help with administrative issues. Staff, professors, and researchers at the University of Victoria and HEALNet for their help and knowledge. And last, but not least, my family and friends who have supported me during all these years. Jennifer Gingell for her encouragement, love, and understanding. My parents, Salmina and Arnold Kagolovsky and brother and sister, Sergey and Anna, for their love, support, and encouragement.

Dedication

For my parents, Salmina and Arnold Kagolovsky,

Who always believe, encourage, and support me,

With Love and Gratitude.

CHAPTER 1: INTRODUCTION

Motivation

With the birth and development of the Internet, large amounts of data and information[1] have become available in an electronic format. Progress in communication and computer technologies has also introduced the notion of a "digital library" (Mannoni 1996; Schatz and Chen 1999). In different implementations of such libraries, many resources are available in multimedia formats, including images, sound, and movies. At the same time, textual and numerical representations are still the most prevalent formats in electronic repositories of data and information. Croft (Croft 1993) writes that "text will continue to be critical due to its unique role as a medium for communication."

While rapid and inexpensive access to electronic data sources can be a great benefit, users can also be overwhelmed and frustrated with a high volume of irrelevant sources when they attempt to navigate "The Information Superhighway." Thus, information retrieval (IR) systems (systems providing retrieval of any type of information from a computer) are critical to realizing the benefits of increased availability of electronic data sources (Salton and McGill

[1] Although it is difficult to give precise definitions of "data" and "information," I am using these terms as they are defined by van Bemmel and Musen in their "Handbook on Medical Informatics" (van Bemmel and Musen 1997). I use the term "data" in relation to uninterpreted items, such as a set of codes or symbols. The term "information" is used when referring to a set of data elements organized in a form that has a meaning. Therefore, the term "information" implies a human being who receives data elements and creates a meaning. Created information sources can be transferred to and analyzed by other human beings.

1983; Kantor 1994; Hersh 1996). IR systems that can better satisfy users' information needs and increase the relevance of the retrieved data and information sources to them are urgently needed. An up-to-date evaluation methodology is crucial to the design of better IR systems (Harter and Hert 1997).

Although research in IR has been going on for the last 50 years, there are still many problems. For instance, there are many factors, other than a search engine, that can contribute to the success or failure of an IR process (Meadow 1992; Hersh 1996; Tague-Sutcliffe 1996; Harter and Hert 1997). Users of a search engine and the IR environment play very important roles. User satisfaction with IR is often based not only on finding relevant data sources, but also on the usefulness of these sources for a problem-solving process. Therefore, there is a need for a shift from only improving the functionality of search engines to a more holistic, systemic view of IR (Meadow 1992; Lancaster and Warner 1993; Hersh 1996). This holistic view requires a different evaluation methodology.

The long history of IR research has resulted in a variety of evaluation approaches. These approaches often characterize IR from a specific point of view. There is a need to compare different approaches toward the improvement of IR. Creation of a theoretical framework that will introduce a common terminology, a basis for a planning of experiments, and a results comparison, will be of great value. I argue that a systems analytic approach to IR evaluation

and the structured model built using this approach can improve the evaluation process. In particular, use of this model can solve long-existing problems of relevance-based evaluation methods that are widely used in IR experiments by some researchers, and intensively criticized by others.

The concept of "relevance" is considered a central concept of information retrieval. In spite of many years of research and discussion there is little agreement on the meaning and application of this concept. At the same time, researchers have problems analyzing and synthesizing results of IR experiments using relevance-based evaluation measures of recall and precision. I argue that an application of the systems approach and new proposed models of IR can open horizons in understanding and usage of the concept of "relevance."

Research issues

This thesis deals with the important issue of improving evaluation of the information retrieval process and its components. I have identified the following problems as having the highest priority:

1. Information retrieval is a complex process and its comprehensive evaluation requires a variety of approaches.
2. There are a variety of evaluation methods with different foci, data gathering strategies, procedures, and means of analysis.

3. Terminology is not well defined, and used inconsistently. It would be desirable to devise a more precise terminology.
4. As a result of the above, there are difficulties in the planning, analysis, and results comparison of IR experiments.
5. Evaluation methods have to focus on the main goal of IR -- users' information needs satisfaction.

Limitations

Some of the boundaries of my research have to be stated. Although there are references in this thesis to the research related to improving functionality and evaluation of the search engines used currently on the Internet, these topics are not discussed in detail. In addition, my research is primarily focused on solving general problems of IR evaluation, rather than specific problems related to IR in health care. At the same time, the results of my research can be used to improve results of information retrieval in any area of research, including health care.

Research goals

This thesis pursues the following goals:

1. Introduction of a common terminology for information retrieval.
2. Creation of a basis for a theoretical framework, including information retrieval models, for analysis of the IR process.
3. Critical analysis of relevance based methods of evaluation, recall and precision.
4. Detailed research of a concept of "relevance" using the systems analytic approach and proposed models of IR.
5. Proposal of a methodology for planning, analysis, and results comparison of IR experiments.
6. Proposal of an improved methodology for evaluating search engines.

Approach and methods

To address the identified research questions the following steps were followed:

1. Conducting a literature review.
2. Critiquing the findings.
3. Identifying problems.
4. Applying the systems analytic approach.
5. Devising an experimental approach for future studies.

A literature review was conducted based on a number of information resources. *Current Contents* was searched in electronic form for the period starting from 1990 for journal articles dealing with information retrieval, IR evaluation, relevance, and cognitive aspects of IR. The *Journal of the American Society for Information Science (JASIS)*, *Information Processing and Management (IPM)*, *Bulletin of the Medical Library Association (BMLA)*, and *Journal of the American Informatics Association (JAMIA)* were browsed for relevant articles. *Annual Reviews of Information Science and Technology* and Proceedings of the Conferences of the *Special Interest Group on Information Retrieval (SIGIR)* within the *Association for Computing Machinery (ACM)* provided many interesting reviews and presentations. Some basic and specialized information on different issues related to information retrieval in general and to health related information sources in particular have been found in books (Salton and McGill 1983; Meadow 1992; Lancaster and Warner 1993; Hersh 1996).

A systems analytic approach was applied to the data gathered during the literature review. To create a structured model of IR the following steps were taken. First, the process of IR was analyzed based on the literature review. This analysis permitted me to identify components of IR and the relationships between them. The IR components were then tested for their "independence" to identify their boundaries. For example, are either a computer-user interface or a

document set parts of a search engine or independent from it? The criterion for the decision was the following: if one version of a component X can be exchanged for another version of the same component, and this exchange does not require changes in other components of the IR process, this component X can be considered an independent one. Therefore, a computer-user interface and a document set are independent from a search engine, as new versions of these components can work with the same search engine. For example, different databases of the National Library of Medicine can work with the same search engine, GratefulMed. The same computer-user interface can be used with different search engines; as well, different interfaces can be tested with the same search engine. Finally, a graphical representation of the structured model was created. The second proposed model is a formal representation of the IR process.

Both models are used for creating a new definition of information retrieval, introducing a common terminology, and analyzing a concept of "relevance" and relevance-based measures of recall and precision. The same models are also applied in proposing new methods for a comprehensive evaluation of search engines and approaches to solving problems associated with a concept of "relevance."

Summary of the subsequent chapters

This thesis consists of the following sections:

1. Literature review and critique of the findings (chapters 2, 3, and 4).
2. Presentation of the proposed conceptual IR framework (chapter 5).
3. Implications of this framework, including a proposal for solutions to the identified problems (chapters 6 and 7).
4. Conclusion (chapter 8).

The results of the literature review are presented in chapters 2, 3, and 4. Chapter 2 gives an overview of information retrieval. It starts with an analysis of IR terminology with the main focus on existing definitions of "information retrieval" and "IR systems." After this, a short history of information retrieval is presented. Chapter 3 presents an overview of IR evaluation. At the beginning, literature about general aspects of evaluation as an activity is discussed. After this, an overview of current IR evaluation approaches is given. This includes traditional evaluation methodologies, such as Cranfield 1 and 2, and their history, theoretical basis, importance, and application in large scale experiments (TREC and STAIRS). This chapter contains a review of a concept of "relevance" and its application in IR evaluation. In addition, relevance-based measures of recall and precision, often considered as the "gold standard" of IR evaluation, are presented, as well as some problems associated with the Cranfield model in

general and relevance-based measures in particular. The chapter 3 also contains a review of alternatives to the Cranfield model of IR evaluation. Chapter 4 presents some of the existing methods of applying the systems analytic approach in information retrieval. Strengths and weaknesses of these models are discussed.

Chapter 5 presents the proposed conceptual IR framework. It includes an overview of the IR process and two models of this process. One of the models presents components of the IR process and relationships between them. The second one models the IR process.

Some of the implications of the proposed framework are presented in chapters 6 and 7. Chapter 6 gives a new definition of information retrieval and demonstrates how the proposed framework helps in introducing a common terminology, as well as in explaining problems associated with a concept of "relevance" and measures of recall and precision. Chapter 7 presents some alternatives for evaluating search engines and "relevance" relationships. The proposed methods are compared with some of the existing evaluation approaches.

The thesis concludes with chapter 8. This chapter gives a summary of the thesis, identifies contributions made, and areas of future research.

CHAPTER 2: AN OVERVIEW OF INFORMATION RETRIEVAL

Terminology

Consistent terminology in a field of study is crucial to the ability to discuss problems constructively. Loose definitions and inconsistent use of terminology have been identified as one of the problems in information science (Schamber 1994). In particular, this problem affects discussions of the fundamental concept of information science: relevance (Froehlich 1994; Schamber 1994). Unfortunately, an agreement on the consistent use of terminology is difficult to achieve.

One of the reasons for this difficulty is related to the basic problem of semantics. The relationship between a real world entity, mental constructs (concepts) about this entity, and associated terminology is usually represented using the "Meaning Triangle" (Ogden and Richards 1946; Scherrer 1998). This theoretical construct demonstrates that the only connection between real world entities and the related terminology occurs through mental constructs. This connection explains the ambiguity with which human beings use terminology. One of the examples of this ambiguity is that people often use the same terms when referring to different concepts and entities. This is called "polysemy." Another variant, called "synonymy," is the usage of different terms to identify similar entities and cognitive structures.

Another reason for terminological difficulties is related to the complexity of a field of study. As information retrieval is a very complex field, combining expertise from computer science, engineering, cognitive psychology, library science, information science, and other disciplines, terminology and definitions used emphasize different aspects of IR (Salton and McGill 1983; Belkin 1984; Meadow 1992; Lancaster and Warner 1993; Hersh 1996). Therefore, the introduction of a consistent terminology and definitions will improve communication between researchers in the field.

To demonstrate terminological difficulties, I will present a variety of meanings associated with definitions and applications of the terms "information retrieval" and "information retrieval system." This presentation will illustrate the need for using consistent terminology and definitions in the field of information retrieval.

Defining "Information Retrieval"

The term "information retrieval" was introduced by Calvin Mooers in 1951 (Cleverdon 1991). However, Mooers' report was not widely circulated and did not make any significant impact. Cleverdon (Cleverdon 1991) points out that his own report about the experiments with the Uniterm indexing language (Cleverdon and Thorne 1954) popularized the term "information retrieval" to the world. Although the field of information science has existed for at least 50 years, I was not able to find a common definition of "information retrieval." Instead,

there are a variety of definitions, which characterize IR from different perspectives, depending on the research background of the authors that propose these definitions. In general, information retrieval is defined either as a field of study, or in a manner based on its structural and functional characteristics.

The definition of IR as a field of study shows relationships between information retrieval and other sciences. Hersh defines IR as "a field at the intersection of information science and computer science, which concerns itself with the indexing and retrieval of information from heterogeneous textual databases" (Hersh 1996, p.3). This definition also specifies two activities that are performed on information (indexing and retrieval) and a type of information (textual). Harter and Hert note that IR as a field of study has practical as well as theoretical characteristics: "IR is an applied field as well as a theoretical one, in that IR systems are designed to help solve real human problems, and IR systems do exist in the real world" (Harter and Hert 1997, p.5).

Structural and functional characteristics of information retrieval are the basis of definitions given by different authors (Salton and McGill 1983; Croft 1993; Hersh 1996; Tague-Sutcliffe 1996). Salton and McGill (Salton and McGill 1983) support Minker's (Minker 1977) emphasis on the functional characteristics of IR. These authors state that information retrieval deals with representation, storage, organization, and access to information resources. These resources can be either

documents or their representations (so-called document surrogates). Another structural-functional definition 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" (Tague-Sutcliffe 1996, p.1). This definition emphasizes the dynamic character of information retrieval, identifies its functions, and recognizes users and their information needs as major components. The tool performing the search, the technical IR system, an interface, and a setting for the process are, however, not specified in this definition.

In many descriptions of information retrieval, additional aspects are identified: the roles of users, intermediaries, and database producers (Salton and McGill 1983; Meadow 1992; Hersh 1996), cognitive processes involved (Belkin, Oddy et al. 1982; Belkin 1984), characteristics of an environment (Lancaster and Warner 1993), the iterative character of IR (Croft 1993), and others. Practically all researchers currently agree that information retrieval has as its main goal users' information needs satisfaction. This aspect of IR is nicely summarized by Harter and Hert who define IR as "a practical act, conducted by a user for a reason - to attempt to satisfy a human need by consulting an information store" (Harter and Hert 1997, p.4). Lancaster and Werner (Lancaster and Warner 1993), however, have a different view on IR:

> "Information retrieval is the major activity engaged in by information centers. The information center, or service, includes libraries, producers of published databases in printed or electronic form, and any other type of service that provides information resources to a population of users." (Lancaster and Warner 1993, p.1)

Compared to the majority of researchers, this view represents a much broader understanding of IR. Lancaster and Werner identify the major function of an information service to be "an interface between a particular population of users and the universe of information resources in printed or other form" (Lancaster and Warner 1993, p.4). To accomplish this function, any information center performs three main activities: "the acquisition and storage of documents; the organization and control of these documents; and the distribution of these documents, or information about them, to users by circulation, literature searches, photocopying, and other services" (Lancaster and Warner 1993, p.4).

Lancaster and Werner identify document delivery and information retrieval as the two major functions through which information centers satisfy users' information needs. Document delivery is considered to be the ability of the center to supply known information resources when needed, while information retrieval is defined as "the ability of the center to retrieve documents on a particular subject, or to provide the answer to a specific question" (Lancaster and Warner 1993, p.6). The authors also specify "information retrieval" and "literature searching" as basically synonymous terms, because both these activities involve searching through a document set to find sources that satisfy an

information request. Meadow (Meadow 1992) presents a similar, comprehensive view of information retrieval as a part of information flow in the world. His view, like that of Lancaster and Werner, extends our understanding of information retrieval as a function common to all information services:

> "IR involves finding some desired information in a store of information or database. Implicit in this view is the concept of selectivity; to exercise selectivity usually requires that a price be paid in effort, time, money, or all three... A library is the best example of an institution devoted to selective retrieval. One does not go there to read the entire collection. One goes to look for something selectively, often something that will satisfy a set of highly individualized information needs... As a practical matter IR is usually implied as a computer activity... While the selection of records from an inventory file or a file of bank depositor accounts can be considered IR, the term is more commonly applied to files of text records or records descriptive of text" (Meadow 1992, p.2-3).

The term "information retrieval" is often restricted to the retrieval of data and information from textual databases. This is due to a history of research in the field that has primarily used textual data and information sources. However, Salton and McGill (Salton and McGill 1983) argue that, theoretically, there is no restriction on the type of items handled in information retrieval. This statement is supported by other researchers who either use the term "document" in its widest sense (Lancaster and Warner 1993) or for whom IR "implies retrieving information of any type from a computer" (Hersh 1996, p.3). This is also consistent with current Internet experience (Mannoni 1996; Schatz and Chen 1999). I concur with this view. If I use the term "document set", I use it in a

manner consistent with Salton and McGill (Salton and McGill 1983), as including text documents, images, sound, and movies.

Although there is no unified definition of information retrieval, the majority of the found definitions emphasize the following characteristics of IR:

1. That IR is a complex process consisting of a variety of components that have structural and functional characteristics, and that interact with other components.
2. That IR is a dynamic iterative process.
3. That IR is related to users' information needs satisfaction.
4. That IR always happens in some kind of environment.

All these characteristics have to be reflected in a comprehensive definition of information retrieval.

What is an IR system?

a) Definitions

Most often, an IR system is characterized as "a unique type of computer application" (Hersh 1996, p.3). IR systems differ from other systems in that they provide information in response to users' queries (Tague, Salminen et al. 1991). More detailed definitions also emphasize other aspects of IR systems, for example, their selectivity. Harter and Hert (Harter and Hert 1997) characterize an IR system as a filter that selects only a small number of documents from a large information store. This allows a user to deal with a manageable number of documents. Although some researchers identify different forms of information retrieval systems -- question-answering systems, data retrieval systems, and text retrieval systems (Lancaster and Warner 1993) -- the majority consider the last kind when discussing IR systems. For example, Harter and Hert (Harter and Hert 1997) refer to an IR system as one that retrieves documents, or references to them, rather than data. Comparisons have been made between IR systems and other types of information systems, for example, database management systems (DBMS), management information systems (MIS), and expert systems (Salton and McGill 1983; Hersh 1996). However, as a result of the rapid development of the computer industry and the growing functionality of IR systems, the differences between these systems are getting less and less clear (Hersh 1996).

Using a broader understanding of IR, Lancaster and Warner (Lancaster and Warner 1993) argue that "information retrieval" can be considered a synonym to "literature searching," as both of these activities deal with the process of searching a documents collection. Based on this, the authors also advocate a broader view of IR systems as any system created to improve literature searching. They cite a library subject catalog and printed indexes as two examples. This view is consistent with that of Meadow (Meadow 1992), who argues that computerized and non-mechanical IR systems use essentially similar principles. As a result, some researchers argue that a library as an institution can be considered as a large and complex IR system (Lancaster and Warner 1993; Harter and Hert 1997)

There are two important types of algorithms that IR systems use: indexing and retrieval (Croft 1993; Hersh 1996). Creation of a representation of a document (indexing) is a very important step for successful retrieval. Salton and McGill (Salton and McGill 1983) demonstrate a new angle by describing information retrieval systems as "consisting of a set of information items (DOCS), a set of requests (REQS), and some mechanism (SIMILAR), for determining which, if any of the information items meets the requirements of the requests." Hersh (Hersh 1996) describes an IR system as consisting of three components: a database, computer hardware and software. Hardware serves as storage for the database, and software processes users' queries and retrieves documents from the

database. A more detailed description has been given by Tague, Salminen and McClellan (Tague, Salminen et al. 1991) who identify such components of a modern IR system as a database of original documents and surrogates, different indexes, and links within and between documents. This system allows different combinations of indexing terms connected by Boolean operators in order to select, rank, and navigate documents and their representations. At the same time, Lancaster and Warner (Lancaster and Warner 1993) identify six major subsystems of an IR system: the document selection subsystem, the indexing subsystem, the vocabulary subsystem, the searching subsystem, the subsystem of interaction between a user and a system (user-system interface), and the subsystem that matches document representations against request representations.

b) Boundaries

According to Harter and Hert, "in its simplest form, an IR system can be viewed as a "black box" that accepts input and produces output" (Harter and Hert 1997, p.4). This is consistent with the view of Robertson and Hancock-Beaulieu (Robertson and Hancock-Beaulieu 1992), who advocate the necessity of an explicit identification of IR system boundaries. At the same time, there are many subsystems inside this "black box". An explicit identification of IR system

boundaries can help in conducting evaluations of different aspects of information retrieval (Robertson and Hancock-Beaulieu 1992)

The main problem is too broad an understanding of the IR system. For example, a database, or a set of documents, is viewed by some researchers as a part of an IR system (Tague, Salminen et al. 1991; Meadow 1992; Hersh 1996). Moreover, researchers often don't explain how boundaries of the system are defined. At the same time, Meadow (Meadow 1992) argues that researchers have to take into account the context of an evaluation situation before making a decision about inclusion or exclusion of a database into IR system boundaries. Inclusion of a document set into boundaries of an IR system without an explicit discussion of this decision results in such phrases as: "searching information retrieval systems is a highly interactive, iterative process" (Borgman, Hirsh et al. 1996, p.568). However, an IR system is itself generally viewed as a searching system. This system searches through a document set to find relevant documents. It is very difficult to agree that an IR system can itself be searched.

To further complicate this issue, some researchers (Buckland and Plaunt 1994) do not consider the user to be a part of the system, while others (Fidel and Soergel 1983) do. Although a user-computer interface is often also considered as part of an IR system, it is possible to evaluate an interface independently from an IR system (Shneiderman 1992). At the same time, some authors believe that an IR

system can have much wider boundaries and, for example, consider a library as a kind of IR system (Lancaster and Warner 1993, p.17; Harter and Hert 1997).

Summary

Although different definitions of "information retrieval" exist, my review of the literature demonstrates that the terminology is not used consistently. There is a need for a definition of "information retrieval" as an activity that will identify components of IR, their functions and interactions. Definitions of IR, as well as its components, have to identify their boundaries clearly to prevent misuse of terms and misunderstandings based on their use. Better terminology will provide a common ground for improved communications between researchers, resulting in improved experiment analysis, planning, and discussion of results. Lancaster and Werner give a very valuable critique of the terminology used in IR research:

> "Clearly, "information retrieval" is not a particularly satisfactory term to describe the type of activity to which it is usually applied. An information retrieval system does not retrieve information. Indeed, information is intangible; it is not possible to see, hear, or feel it. We are "informed" if our state of knowledge on a subject is somehow changed. Giving a requester a document or a reference to a document on lasers does not inform him or her on the subject of lasers. Information transfer can take place only if the user reads the document and understands it. Information, then, is something that changes a person's state of knowledge on a subject. This may not be a very precise definition, but it is the best that we can offer" (Lancaster and Warner 1993, p.11).

It is clear that while information systems can provide a user only with data and information sources, information transfer is related to a user's assimilation of

these sources. This assimilation consists of extracting meaning from retrieved data and information sources, and introducing changes in a user's state of knowledge. Terminology in the field of IR has to reflect our current understanding of cognitive processes involved in information retrieval, rather than clinging to old definitions reflecting our views during earlier stages of the IR research.

To solve problems related to defining an IR system, two approaches have been proposed and used. Both of these approaches attempt a more precise usage of terminology. The first approach focuses on the introduction of a more precise terminology. Thus, some of the researchers use special terms for a computer system performing indexing and retrieval of data and information sources. For example, Robertson and Hancock-Beaulieu (Robertson and Hancock-Beaulieu 1992) advocate the use of the term "mechanism" for a computer system, while Borgman, Hirsh and Hiller (Borgman, Hirsh et al. 1996) use the term "query-matching system." In addition, the development of the Internet has introduced the term "search engine." The second approach is directed toward a more precise definition of terms. Robertson and Hancock-Beaulieu (Robertson and Hancock-Beaulieu 1992) present a variety of views on the concept of an IR system in evaluation research. The authors recommend that researchers explicitly identify boundaries of a system under investigation, including an IR system.

Historical overview of information retrieval

To reveal the origins of current views and problems regarding the main concepts of IR, it is useful to look at changes in research focus in the field over time and how the changes have affected our understanding of information retrieval. I will first outline historical changes in IR in general, and then evaluation approaches in particular.

I argue that the most important directions of the research in the field of IR have been improvements to technical information retrieval systems (search engines), attention to users and their characteristics, and understanding the dynamic character of the IR process with its important human-computer interaction. Robertson and Hancock-Beaulieu (1992) identify the three major changes in our perception of information retrieval as:

1. research into the concept of "relevance,"
2. cognitive aspects of user information seeking behavior, and
3. the increasingly interactive character of information retrieval.

The authors argue that these changes "have profoundly influenced the nature of the evaluation problem" (Robertson and Hancock-Beaulieu 1992, p.459).

Two major factors contributed to the beginning of information retrieval as a field of study in the 1950s (Cleverdon 1991). First, many documents that were not available to a wide community during WWII were released. These documents

were in a various format and their number was overwhelming according to the standards of that time. There was a big demand for effective methods of indexing and finding required documents among the large number available. The second factor was related to intensive work on computers that were considered a perfect tool for organizing, indexing and retrieving documents.

In the beginning, IR research was concerned with the improvement of technical IR systems[2] (Salton and McGill 1983; Cleverdon 1991; Salton 1992). The main focus was on improving computer algorithms in order to better handle large quantities of electronic data and information resources. Computers were expensive, difficult to operate, and not very user friendly. They were in the hands of engineers, and potential users did not interact with computers directly. Queries were submitted to the intermediate person, searches were performed in batches, and answers could take days (Cleverdon 1991; Lancaster and Warner 1993; Hersh 1996). In the 1970s information systems were still not powerful enough to store large databases, and were only able to work with bibliographic databases (Lancaster and Warner 1993; Hersh 1996). As a result, research was focused primarily on development and improvement of techniques for storing and retrieving text documents (Salton and McGill 1983). At the same time, even during this highly computer-oriented period, some researchers realized a need

[2] I use the terms "technical IR system" and "search engine" interchangeably.

for a wider view of IR (Wilson 1973; Swanson 1977; Fidel and Soergel 1983; Belkin 1984; Meadow 1986).

This wider view started to form during a second period, when the role of a user was understood as a very important one. Salton and McGill wrote that although "most practitioners interested in the design and operations of actual retrieval systems are concerned only about applied computer science," information retrieval as a science has two links: one to computer science and another one to "behavioral science, since retrieval systems are designed to aid human activities" (Salton and McGill 1983, p.xii). The interest in exploring the link between information retrieval and behavioral science was triggered by a discussion of the concept of "relevance" in IR (Cuadra and Katter 1967; Cuadra and Katter 1967; Lesk and Salton 1969; Saracevic 1970; Harter 1971; Wilson 1973; Saracevic 1975; Meadow 1985; Swanson 1986; Eisenberg and Schamber 1988; Schamber, Eisenberg et al. 1990). This concept is considered a central one in the field of information science (Saracevic 1975; Harter 1992; Froehlich 1994; Hersh 1994; Schamber 1994; Mizzaro 1997). A connection between this concept and user satisfaction with IR results is widely accepted by a majority of researchers. However, the meaning of "relevance" and appropriate methods for its evaluation continue to be debated. As users play an important role in evaluating "relevance," their characteristics are intensively researched and discussed (Fidel and Soergel 1983; Fidel and Crandall 1997). One of the most important concepts

related to cognitive processes involved in the information retrieval process is a concept of an "anomalous state of knowledge" (ASK). It was introduced by Belkin (Belkin, Oddy et al. 1982; Belkin 1984) and is considered a central concept in understanding the role of a user in the information retrieval process (Robertson and Hancock-Beaulieu 1992; Schamber 1994).

A third big change is related to a realization of the interactive character of IR (Robertson and Hancock-Beaulieu 1992). Within the last 15 years users have started to interact directly with information systems and perform more searches by themselves rather than through librarians. Information retrieval has always been a dynamic process with a user examining results and taking some actions, for example, re-evaluating information needs, or changing a request. However, real-time user-computer interaction has not always been possible. More powerful personal computers, improvement of the user-computer interface, easier connection to electronic information sources, and development of the Internet and search engines have made this possible (Hersh 1996). Tague-Sutcliffe (Tague-Sutcliffe 1992) writes that since her review in 1981, the field of information retrieval has experienced tremendous technological changes that have resulted in a shift in the paradigm, identification of new problems, and new research areas:

> "Technologies such as CD-ROM and improved communication networks have widened the availability of computer-based retrieval systems. Others, such as full-text databases and hypertext and hypermedia

> systems, have enlarged our notion of what constitutes an information record in an information retrieval system. A paradigmatic shift has occurred in the research front, to user-centered from system-centered models." (Tague-Sutcliffe 1992, p.467)

As a result of the outlined changes, we now have a wider and a more detailed view of IR. This view is not limited only to a technical IR system, but includes users interacting with this system and the environment in which the IR process occurs. At the same time, our image is more detailed since many researchers investigate different components and parameters of information retrieval. Based on the changing view of IR, new and more sophisticated methods of evaluation have been introduced. A review of evaluation methods, in relation to the outlined changes in our view of information retrieval, is needed for a better understanding of evaluation problems.

CHAPTER 3: AN OVERVIEW OF IR EVALUATION

An Overview of Evaluation in General

Even a superficial search for literature sources discussing evaluation reveals a large number of publications and the complexity of the problem. It leaves one with a clear impression that evaluation is a difficult and often confusing research activity, as "there can be no single solution to the problem of evaluation. There is, instead, an interdisciplinary field of evaluation with an extensive methodological literature" (Friedman and Wyatt 1997, p.18). Based on the reviewed literature, evaluation can be considered as a general research activity used in many fields of study. Issues related to the general aspects of evaluation continue to be the focus of many publications (House 1980; Guba and Lincoln 1981; Rossi and Freeman 1989). Currently, more and more research is devoted to evaluation in health informatics (Anderson, Aydin et al. 1994; Friedman and Wyatt 1997).

The number of sources that discuss evaluation aspects of information retrieval is so high that even reviews of this literature have to concentrate on specific areas. For example, Harter and Hert (Harter and Hert 1997) discuss only the evaluation of search engines; Shaw (Shaw 1991) and Vickery and Vickery (Vickery and Vickery 1993) review the design and evaluation of the search interface; Kantor (Kantor 1994) considers the role of experimental techniques in IR; Mizzaro

(Mizzaro 1997) reviews the history of "relevance," and Schamber (Schamber 1994) gives an overview of the literature concerning a connection between the concept of "relevance" and the information behavior of users. Several reviews of psychological, cognitive, and other user-oriented research related to IR have also been published (Daniels 1986; Allen 1991; Sugar 1995; Hert 1997). In this chapter I will present a short overview of the main issues related to evaluation in general and IR evaluation in particular. Some of these issues will be discussed in more detail later in this thesis.

What is evaluation?

Although the term "evaluation" is used often in the literature, it has no single accepted definition. Usually, this term describes a wide variety of data collection activities designed to answer questions and assist in the decision-making process: "Most people understand the term "evaluation" to mean measuring or describing something, usually to answer questions or help make decisions" (Friedman and Wyatt 1997, p.1). In the IR evaluation field, researchers use similar definitions of "evaluation." For example, Harter and Hert (Harter and Hert 1997) adapt a definition of evaluation by Hernon and McClure (Hernon and McClure 1990) that states that "the process of identifying and collecting data about specific services or activities, establishing criteria by which their success can be assessed, and determining both the quality of the service or activity and

the degree to which the service or activity accomplishes stated goals and objectives" (Harter and Hert 1997, p.5).

The reasons for evaluation

There are different reasons why evaluation of information retrieval is important. Some of these reasons are:

1. Information retrieval can assist users in the decision-making process. If we want to improve decision-making, there is a need to evaluate the impact of changes that are made to any component of IR.
2. We would like to know if information systems help the users to do what they want to do -- if they satisfy users' information needs.
3. Installation of information systems in organizations, user training, and support of the systems require a considerable investment of time and money. There is a need to evaluate whether this investment produces the expected impact, for example, if information systems improve processes and their outcomes for organizations.
4. Organizations working on improving search engines need effective methods for evaluating changes made to algorithms and user interface.
5. Providers of information resources require information about usage of these resources by users.

Goals of evaluation

Tague-Sutcliffe (Tague-Sutcliffe 1996) specifies user satisfaction as the main concern of an IR system evaluation. She does not mean only individual users, but all actual and potential users in the community. Tague-Sutcliffe states that the purpose of evaluation is "to lead to improvements in the information retrieval process, both at a particular installation and more generally" (Tague-Sutcliffe 1996, p.1). She argues that to achieve general improvements there is a need to address theoretical issues and make comparisons under controlled experimental conditions. Such controlled conditions can help to distinguish between general improvement achieved versus some specifics of local installation.

Other goals of evaluation depend on the research interests of scholars. For example, researchers that are more interested in improving functionality of search engines argue that the main goal of evaluation is "to evaluate dispassionately the power of modern indexing and search methodologies" (Salton 1992, p.448). Some other foci of IR evaluation research are cognitive processes of users, human-computer interface, and characteristics of databases.

Types of IR evaluation

There is a wide range of evaluation questions, ranging from technical characteristics of specific systems to their effects on people and organizations.

However, researchers have to select the best or most appropriate set of questions to explore a particular situation. Based on the kinds of questions that are seen as the most important, researchers choose appropriate evaluation approaches and data collection methods. Friedman and Wyatt note that methods of evaluation depend on "what is being evaluated and how important the decision is" (Friedman and Wyatt 1997, p.1).

There are different classifications of IR evaluation. For example, Meadow (Meadow 1992) divides IR measures into two broad categories: evaluation of performance and evaluation of outcome. Measures of performance are descriptive of what happens during the use of an IR system. Measures of outcome are descriptive of the results obtained. Hersh (Hersh 1996) specifies two types IR system evaluation: macro-evaluation and micro-evaluation. Macro-evaluation (also viewed as clinical or field evaluation) is an outcome-oriented type of evaluation that investigates the IR system as a whole and its overall benefit. Micro-evaluations, usually performed in a controlled setting such as a laboratory, serve to assess different components of the system and their impact on performance. Lancaster and Warner (Lancaster and Warner 1993) define three levels of evaluation:

1. Effectiveness of the system and user interaction 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 measures the unit costs of various aspects of the retrieval system.
3. Cost-benefit. This 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.

Harter and Hert (Harter and Hert 1997) distinguish between the following types of evaluation:

1. Information evaluation and information system evaluation.
2. Basic evaluation research.
3. Evaluation of operational IR systems and subsystems.
4. Testing and evaluation of specific IR subsystems, such as indexing, experimental processing techniques and algorithms, and relevance feedback.
5. Evaluation of settings, for example, library or teaching environments.
6. Evaluation of libraries and library subsystems.

The variety of evaluation problems in information retrieval, as well as different research methodologies available, present a challenge for scholars in choosing the most appropriate evaluation approaches for the specific problem. Researchers in IR have to be aware of different qualitative and quantitative evaluation methods, and ready to explore new paths to achieve better IR performance. Robertson and Hancock-Beaulieu (Robertson and Hancock-

Beaulieu 1992) argue that design of better IR systems requires more effective evaluation methods. "The challenge for the next decade," they write, "is to explore the multiple dimensions and components of the new generation of information retrieval systems by experimenting with a diversity of evaluative approaches" (Robertson and Hancock-Beaulieu 1992, p.465). At the same time, practically all researchers working with evaluation methodologies caution against the too optimistic approach that evaluation can solve all problems (Friedman and Wyatt 1997). There are many potential evaluation questions. Different groups can disagree on what constitutes the best set of evaluation questions. At the same time, there are different ways to explore evaluation problems, and each evaluation methodology has its pros and cons. Probably the most important advice is that "there is no such thing as a perfect study" (Friedman and Wyatt 1997, p.19).

Problems in IR evaluation

Although the evaluation of any information system is difficult (Friedman and Wyatt 1997), evaluation of IR can present some additional problems. These problems are primarily associated with the complexity of information retrieval. The varieties of technical, human and organizational factors that can affect IR are the main components of this complexity. In 1970, one of the pioneers of IR evaluation, Cyril Cleverdon, wrote: "As of now, it is probably true to say that all information systems represent an act of faith on the part of someone or some

organization" (Cleverdon 1970). Although considerable research efforts have been made since then, IR evaluation still presents tremendous challenges for researchers. Salton (Salton 1992) writes that some experts doubt the validity of IR evaluation methodologies and available experimental results, and question the integrity of the field as a whole.

The evaluation process often involves some kind of measurement. At the same time, Meadow (Meadow 1992) rightly argues that in IR, as in many activities involving humans, there are no equivalents to the physical measures often used in the sciences. Researchers often have problems with defining measures used, as well as in choosing an appropriate method of measurement. There is a need to define as precisely as possible what is measured/evaluated. For example, many researchers claim that they evaluate information retrieval systems. However, experimenters often use different definitions or boundaries of the system being evaluated. This creates problems in the analysis of results and their comparison to results of other researchers.

Tague-Sutcliffe (Tague-Sutcliffe 1996) argues that most of the major problems and issues in evaluation of IR exist as the result of a lack of agreement about characteristics of the IR process and its components. These are the questions that she considers the most important:

1. Who should be used in IR experiments for making relevance judgements: real information needs or subject experts?
2. Should information retrieval evaluation involve actual retrieval processes?
3. How can we balance evaluation of individual components of the retrieval process in addition to the process as a whole?
4. What kind of aggregation is appropriate for the measures used in evaluating IR systems?
5. What is the value of an analytic or simulatory approach in comparison to an experimental one?
6. How can interactive retrieval systems be evaluated?
7. What kinds of generalization are possible from IR tests?

In their review, Harter and Hert (Harter and Hert 1997) also identify a number of problems and questions that present a challenge for researchers in information retrieval:

1. What is the meaning of "information retrieval"?
2. What is the meaning of "IR evaluation"?
3. What components of IR should be evaluated?
4. Do we have to include users, their cognition, and work environments in the evaluation process?
5. Should IR subsystems be evaluated?

6. What are the differences between IR systems evaluation in different fields of study (computer science, psychology, library science, and others)?
7. How is the evaluation of operational systems serving real users different from the evaluation of experimental systems?

Robertson and Hancock-Beaulieu (Robertson and Hancock-Beaulieu 1992) present slightly different formulations of the same problems:

1. Laboratory vs. operational system tests.
2. Black-box vs. diagnostic experiments.
3. Qualitative vs. quantitative evaluation methods.
4. Identifying the boundaries of the system(s) to be experimented upon.

Salton (Salton 1992), in a concise and clear description of the problems related to IR evaluation, identifies evaluation measures of retrieval effectiveness, in particular recall and precision, as the focus of many discussions in the literature. In general, these discussions address such problems of retrieval system tests design as relevance judgements and construction of the so-called "recall base". Other researchers object to generalizing results of controlled experiments using small test collections to large scale systems operating in real life environments.

Current IR evaluation approaches

Traditional IR evaluation: The Cranfield model

The origins of IR evaluation are usually attributed to the late 1950's and early 1960's, when the National Science Foundation began sponsoring a program in systems evaluation (Salton 1992; Tague-Sutcliffe 1996). The first major evaluation studies in IR under this program were conducted by Cyril Cleverdon at the College of Aeronautics in Cranfield, England (Cleverdon 1967; Cleverdon 1991). The Cranfield experiments formed the leading paradigm in IR and are often considered as a prototype design for many other studies.

The original Cranfield model is a laboratory model for evaluation and comparison of technical IR systems, or search engines. This model requires having a test collection of documents, a set of queries, and relevance judgements regarding relationships between the documents and the queries. Relevance-based measures of recall and precision are the most important measures used in the Cranfield model.

A good history of the Cranfield experiments was written by Cleverdon himself (Cleverdon 1991). Cleverdon did interesting research in IR even prior to getting involved with Cranfield tests. In his experiments with systems in the Netherlands, Cleverdon came to the conclusion that "for a valid comparison

between systems, it would be necessary to control conditions in such a way that performance could be related to economic factors" (Cleverdon 1991, p.4). This view is consistent with the current macro-evaluation approaches that try to assess the economical impact of search engines on organizations and their functionality (Lancaster and Warner 1993; Hersh 1996).

a) Cranfield 1

In his presentation at the Special Libraries Association Conference in Detroit (1955), Cleverdon proposed that there is a need for an independent evaluation of different indexing languages and techniques. As a result of this presentation, the National Science Foundation (NSF) gave him a grant in 1957 for a comparative evaluation of four systems: a conventional classification (the Universal Decimal Classification), a conventional alphabetical subject index, a purposely devised schedule of facet classification, and the Uniterm System of Coordinate Indexing. This experiment has been referred to as Cranfield 1. Its database consisted of 18,000 papers in the field of aeronautical engineering. The papers were indexed using each of the four systems. Cleverdon also introduced some factors into the experiment to prevent the bias indexers develop as a result of their learning during the indexing process. Experimenters worked with a total of 1,200 queries obtained from several hundred researchers in 58 organizations located mainly in England and the USA. Each question was based on a single document in the test

collection, and a search was considered successful if that particular paper was retrieved. Experiments showed that all systems performed very close to each other and achieved approximately 74-82% success in retrieving the required paper. Another series of experiments also showed that end-users worked equally well with all indexing systems.

This experiment was intensively criticized. The most interesting example being Swanson's critique of the use of search questions based on documents in the test collection (Swanson 1965). In response to this critique, Cleverdon argues that "[t]he validity of this point was somewhat lessened in that neither Swanson nor anyone else had been able to propose any other practical technique which would have overcome so effectively the problem that is associated with the determination of relevance, an aspect of evaluation testing which has still not been satisfactorily settled, in spite of being the subject of dozens of papers" (Cleverdon 1991, p.5). Cleverdon also explains the reason why such a decision was made:

> "The crude method used in Cranfield 1 was a reaction to the debacle of an earlier attempt to compare two systems. In this, two groups each indexed some 15,000 documents and carried out a number of searches in their own system. When they came together to consider the results, there was a total failure to agree on the relevance of the different sets of citations which each group had retrieved, and at the end of the second day of meetings, they were still arguing about the meaning of the first search question" (Cleverdon 1991, p.5).

This shows that intensive discussions dealing with the concept of "relevance" and relevance judgements have taken place since the beginning of IR evaluation research.

b) Cranfield 2

The main goal of Cranfield 1 was an investigation of the impact of index languages on economic factors. Therefore, it attempted to simulate an operational setting. The next step was an attempt to evaluate the performance of index languages in isolation from other factors. This required creating a laboratory type situation as free as possible from operational variables. Considering that all index languages are combinations of recall and precision devices, the objective of Cranfield 2 was to measure the effect of each of these devices, alone or in any possible combination, on recall and precision.

The design of Cranfield 2 was different from Cranfield 1, as it was clear that for such an experiment to be successful there had to be a decision made about the relevance of every document in a dataset to each query used. This requirement essentially imposed a limitation on the size of a collection of documents (dataset). Two hundred authors of recently published research papers were contacted. They were asked to formulate in the form of a question the problem that their papers were addressing, and to add supplementary questions that were

addressed in their research. The authors then were requested to indicate, using a scale from 1 to 5, the level of relevance of cited references to every question addressed in their research. The test collection was finally composed of 1,400 documents. A group of six students screened these documents against 279 questions. Documents considered relevant to particular questions were sent to the originator of the question for a judgement.

The next step was the indexing of the documents in the test collection. The indexer first recognized the concepts in the document. These concepts were sometimes expressed in a single word, but more often in two or three words. A weighting in the range of 1 to 3 was assigned to each concept in a document as an indicator of its relative importance within the document. After this, a list of every single word in the concepts was established for every document, with an appropriate weighting given to each listed word. On average, every document had 33 single terms, with 14 given the top weighting and an additional 8 given the medium weighting. After this, concepts were combined into themes.

Cranfield 2 was trying to compare index languages by introducing various recall and precision devices. Index languages varied from single terms in the natural language of the documents to more complex indexing schemes. The original hypothesis was that introduction of more complex recall and precision devices would inevitably improve the performance. In other terms, an indexing process

that uses thesauri was expected to show a better performance than one using terms from natural language. However, this hypothesis proved to be wrong. The results of Cranfield 2 were similar to ones obtained by Salton (Salton and McGill 1983; Salton 1992). His research consistently showed that indexing using single terms is more efficient than using phrases, and that synonym dictionaries improve performance, but more complex organizations of terms, such as hierarchies, are not as effective as could be expected. All these results were contrary to those expected and stimulated much argument. Again, the majority of these arguments centred around relevance decisions used in these experiments (Harter 1971; Swanson 1971; Harter 1996).

c) Summary of the Cranfield studies

a) The Cranfield studies were trying to evaluate technical IR systems (search engines).

b) Questions were based on individual papers.

c) Questions did not represent users' information needs.

d) For each question, there was a single "core" relevant document, the one that was used to create the question. Additional relevant documents were found by students screening the collection. Relevance judgements were confirmed by the authors of the questions.

d) The importance of the Cranfield experiments

The Cranfield experiments built a foundation for evaluation of information retrieval by:

a) Setting a standard for experimental design in IR evaluation performance.

b) Identifying various factors that can influence IR performance and providing test data that were used for comparison in many later studies.

c) Showing the influence of indexer error on retrieval performance.

d) Using measures of recall and precision, and demonstrating an inverse relationship between them.

e) Starting intensive discussion about the evaluation of information retrieval, and especially about the concept of "relevance."

f) Leading to research efforts that have resulted in improved indexing and retrieval algorithms.

e) Large scale experiments using the Cranfield model

Some of the critiques of the Cranfield experiments relate to the small size of the collections and controlled settings (Meadow 1992; Hersh 1996; Tague-Sutcliffe 1996; Harter and Hert 1997). Experiments using large test collections had been difficult to conduct until the 1980s due to limitations in data storage and high costs associated with using computers.

(1) The STAIRS experiment

One of the first experiments with a large collection of full-text documents was conducted using the STAIRS (STorage And Information Retrieval System) created by IBM (Blair and Maron 1985; Blair and Maron 1990; Blair 1996). This study was an attempt to evaluate a commercial information system with a large dataset covering approximately 350,000 pages of legal materials. In this study, researchers made considerable efforts to ensure that the majority of relevant documents were identified. This was done using the original method of investigating subsets of the database that had a high probability of containing relevant documents. The STAIRS study is also very interesting as it was done in the context of the defense of a $237 million lawsuit. The environment of this experiment involved political, ethical, and legal issues. Results of the study have been intensively discussed in the literature (Blair and Maron 1985; Salton 1986; Blair and Maron 1990; Blair 1996).

(2) The TREC experiments

The Text REtrieval Conference (TREC) is a large IR evaluation project. It is coordinated by the National Institute of Standards and Technology (NIST) and sponsored by the Advanced Research Projects Agency (ARPA) of the United States Department of Defense. TREC participants have gathered once a year

since 1992, and the proceedings of these conferences have been published annually since 1993. At this conference researchers from academia, industry, and government evaluate and compare the performance of their search engines on large test collections (several gigabytes) of full text documents from different areas of knowledge (Harman 1995; Sparck Jones 1995). Test collections include Wall Street Journal, Associated Press newswires, San Jose Mercury News, Ziff-Davis computer articles, Federal Register reports, US Patents, and other sources. TREC experiments use realistic information requests, and relevance judgements.

TREC has two tasks, routing and ad hoc, that simulate a usual retrieval situation. In routing, the assumption is that the same queries are used, but that users are looking for new documents. This task models an execution of a user profile for a selective dissemination of information (SDI). For the routing task, topics and their relevant documents are known. Participants create queries from the topics, train search engines on special training data sets, and use refined queries against testing data sets. In the ad hoc task, which models a regular searching situation, new queries are submitted to the existing data sources. In this task, queries are generated from new information topics. These queries are then submitted to the search engines, which search existing data. In an ad hoc task relevant documents are not known before the test.

Approximately 350 topics have been created by specially assigned researchers, who also make relevance assessments of the retrieved documents from the pool of 100 or 200 documents retrieved by the participating search engines. Comparison between search engines in general, as well as of how they perform for each topic, are made with the relevance-based measures of recall and precision. Unfortunately, variation in the parameters of the participating search engines does not allow for conclusions about the best methods for achieving the best performance (Smeaton 1993; Kantor 1994; Harman 1995; Sparck Jones 1995; Beaulieu, Robertson et al. 1996; Harter 1996; Harter and Hert 1997; Voorhees 1998; Zobel 1998)

Lately, researchers participating in TREC have become interested in evaluating interactive systems (Beaulieu, Robertson et al. 1996; Harter and Hert 1997). This interest has led to establishing an "interactive track" within TREC. This track has introduced a simulation of realistic interactive searching in a controlled experimental environment.

The main critique of TREC experiments is that they use relevance assessments similar to the Cranfield model (Harter and Hert 1997). At the same time, researchers try to make sure that as many relevant documents as possible are found for each topic. The consistency of relevance judgements is constantly tested. However, it has been shown that measures of recall and precision that

use binary relevance judgements underestimate real abilities of search engines (Smeaton 1993; Shaw, Burgin et al. 1997). Other critiques are related to unrealistically long topics and their difference from queries that are usually submitted to search engines by real users.

f) Problems with the Cranfield model

There is an extensive literature discussing problems associated with the Cranfield experimental model (Swanson 1965; Salton and McGill 1983; Swanson 1988; Salton 1992; Ellis 1996; Harter 1996; Harter and Hert 1997). One of the most comprehensive and up-to-date summaries is presented in the review of the evaluation of information retrieval systems by Harter and Hert (Harter and Hert 1997). The authors identified four groups of problems: validity and reliability, generalizability, usefulness, and conceptual issues. Some of these issues are:

a) That a user is omitted from this model.

b) That the model does not take into account user interaction with a system.

c) Problems with measures of recall and precision.

d) That the concept of "relevance" used in this experimental design is incompatible with real-world experience.

e) The use of relevance judgements that do not reflect user information needs.

f) The lack of applicability of results to operational settings.

g) That these experiments do not provide a framework for comprehensive IR evaluation.

Relevance-based measures of recall and precision

Recall and precision are relevance-based evaluation measures of IR that were first proposed by Kent and colleagues (Kent, Berry et al. 1955). These measures have been intensively used since the Cranfield experiments (Cleverdon and Keen 1966). Recall and precision have formed the basis of the evaluation of search engines, and are considered the "gold standard" of IR evaluation by many researchers (Salton and McGill 1983; Cleverdon 1991; Hersh 1996; Harter and Hert 1997). Although many different measures of IR system performance have been proposed, relevance-based measures of recall and precision are still the most common in laboratory, as well as operational settings (Salton and McGill 1983; Salton 1992; Hersh 1996).

Recall is the proportion of all documents in a database judged by a human to be "relevant" to the query retrieved by a search engine, and precision is the proportion of retrieved documents that are "relevant" to the query. To calculate recall and precision, the following four document sets have to be formed and counted for a given query:

1. Relevant and retrieved documents.
2. Relevant and non-retrieved documents.

3. Non-relevant and retrieved documents.
4. Non-relevant and non-retrieved documents.

The evaluation measures of recall and precision were originally proposed and used based on a system view of relevance (Schamber 1994). This view emphasizes a match between the terms of the query and indexing terms that a search engine uses to create a representation of a document. The system view of relevance is also called objective, because it assumes that a query reliably represents a user's information need, and that a relevance judgement can be made by any user. This view uses a binary relevance judgement ("relevant" or "not relevant").

The theory of recall and precision measures has been discussed by many researchers. The inverse relationship between these measures, a phenomenon that was originally theoretically discussed by Robert Fairthorne, was also found experimentally (Cleverdon 1991). Some researchers applied complex mathematical models for analysis of relationships between recall and precision (Raghavan, Jung et al. 1989; Buckland and Gey 1994). Changes in indexing and retrieval algorithms were proposed to improve performance of search engines as result of this research.

The theoretical basis and practical applications of recall and precision were intensively criticized practically from the beginning of their use in Cranfield

experiments (Swanson 1965; Harter 1971; Swanson 1971; Cooper 1976). The critique of recall and precision measures is usually related to

1. The design of IR experiments (for example, who designs them, when they are designed, and how relevance judgements are made)
2. The calculation of relevance-based measures.
3. Use of these measures.
4. The interpretation of results.

For example, the main assumption in using these measures for comparing search engines' performance is that "the average user is interested in retrieving large amounts of relevant materials (producing a high recall performance), while at the same time rejecting a large proportion of the extraneous items (producing high precision)" (Salton 1992, p.442). However, it is well-known that users are not always interested in a truly exhaustive search, which finds everything that might be related to a query (Cleverdon 1991; Hersh 1996). And "non relevant" retrieved documents do often provide important prompts to a real user.

The introduction of user-centred and interactive perspectives on IR has resulted in new views of relevance, such as information and situation views (Schamber 1994). These views take into account the large body of research about relevance judgements and the many factors that can influence them (Saracevic 1975; Schamber, Eisenberg et al. 1990; Harter 1992; Bruce 1994; Froehlich 1994; Hersh 1994). For example, it was shown that recall, in particular, is incompatible with

the utility-theoretic paradigm (Cooper 1976). According to this paradigm, retrieval effectiveness is measured by determining the utility of the retrieved documents to the users. Cooper showed that the utility and the relevance of a document are different: a document can be relevant to the query, but its utility to the user will be zero. This can happen in a case when a user knows about a document before her search. In another case, a single document can be exactly what the user needs, but at this point the recall value will be very low. Cooper argued that often the precision value alone can be enough, or that some other measures can be used instead of recall and precision.

The main critique from the beginning of the Cranfield experiments was related to the use of relevance assessment (Swanson 1965; Swanson 1971; Harter 1996). Salton (Salton and McGill 1983; Salton 1992) 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" (Salton 1992, p.445). 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" (Salton 1992, p.442). Ideally, the results are evaluated by the same person.

Closer scrutiny showed that relevance judgements are subject to many influences that can affect them even during the same experiment (Schamber 1994). Many experiments do not support Salton's arguments (Harter 1996). Hersh (Hersh, Pentecost et al. 1996) asserts that "recall and precision, may have serious problems in their external validity, at least as they are usually measured." Smeaton (Smeaton 1993) 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 problematic. One of the problems is a difference between users' information needs and their expression in queries. The second problem is that information retrieval is an interactive process and users' information needs can change during experiments. The third one is that the results of the TREC experiments disclose the usefulness of a variation in performance characteristics of IR systems that can support different users in a variety of ways (Smeaton 1993). Recall and precision do therefore not suffice to compare IR qualities.

Cleverdon (Cleverdon 1991) also advocates caution in interpreting results of recall and precision measures obtained in experimental settings. He argues that

in operational situations, a high level of recall is not required, as the majority of users are looking only for a few relevant papers. Cleverdon cites results obtained by Lantz (Lantz 1981) who found that the number of relevant citations retrieved in on-line searches was significantly higher than the number of citations used. In regards to precision, even if in an experiment the precision level could be raised from 30 to 40% at the same level of recall (which is definitely a great improvement), an average user in an operational setting would hardly notice this improvement.

Cleverdon (Cleverdon 1991) argues that the ultimate measure of an IR system in operational settings must include cost. He proposes a measure that takes into account an assumption that the objective of an IR system is to retrieve relevant citations without retrieving non-relevant ones, at the lowest possible cost. His argument shows in a very convincing way that such measures can change our perception of IR test results and emphasizes the extreme care necessary in interpreting test results and their transfer from laboratory into operational settings.

The use of recall AND precision has also been criticized. For example, some researchers argued that it would be better to have one evaluation measure rather than two (Voiskunskii 1997). Others suggested that recall and precision do not reflect either the size of a data collection or the number of retrieved documents

(Fairthorne 1964). Many alternative search engine performance evaluation measures have been proposed (Harter and Hert 1997): Swets' E-measure (Swets 1963; Swets 1969), Cooper's expected search length (ESL) (Cooper 1968), and Losee's expected precision measure (Losee 1996; Losee 1997). However, these measures have not been widely used, partially because of difficulties associated with their calculation and interpretation (Salton and McGill 1983; Salton 1992; Harter and Hert 1997). A good summary of the importance of recall and precision was written by Hersh: "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" (Hersh 1996, p.38). Different questions and problems related to relevance-based measures of recall and precision will be addressed later in my thesis. One of the most important issues addressed is: "Do recall and precision really evaluate search engines?"

New Paradigms of Information Retrieval Evaluation

The growing awareness of the limitations of the Cranfield model in general, and of relevance-based measures in particular, has turned attention towards the users of IR systems and computer-user interfaces. Researchers have started to recognize the holistic and dynamic character of information retrieval. A new paradigm for IR, the cognitive paradigm, has been proposed (Dervin and Nilan 1986). This paradigm is also called behavioral, user-centered, and user-oriented (Tague-Sutcliffe 1992; Schamber 1994).

User-centered models of IR (Fidel and Soergel 1983; Meadow 1992; Lancaster and Warner 1993) emphasize different views of the IR process by defining boundaries of IR (Robertson and Hancock-Beaulieu 1992), as well as cognitive and behavioral aspects of information retrieval (Schamber 1994; Ellis 1996). Some of the new models identify components of the IR process (Tague-Sutcliffe 1992), their functions and important variables (Fidel and Soergel 1983), and consider users' multiple interactions with a search engine during the same session (Su 1992; Schamber 1994). It is obvious, then, that user characteristics are important (Fidel and Soergel 1983). Such concepts as "relevance," "cognition," "user behavior," and "interaction" contribute greatly to a "changing view of the boundaries of the system" (Robertson and Hancock-Beaulieu 1992). This has resulted in "a paradigmatic shift ... in the research front, to user-centered from system-centered models" (Tague-Sutcliffe 1992) and to definitions of IR emphasizing cognitive, behavioral and affective aspects of the IR process (Belkin 1984; Wilson 1994; Ellis 1996).

There is a growing interest in assessing the cognitive, behavioral and affective aspects of IR through qualitative methods used in social sciences and humanities (Wilson 1981; Fidel 1993; Ellis 1996). Wilson (Wilson 1981) has proposed to use qualitative research in studying the information needs underlying information seeking behaviors of users. Park (Park 1994) has argued for more research using a qualitative approach in the investigation of users' criteria of relevance.

Schamber (Schamber, Eisenberg et al. 1990) has studied the underlying meaning of users' evaluation criteria in terms of contextual and multimedia information retrieval.

Based on the cognitive approach, Nilan, Peek, and Snyder (Nilan, Peek et al. 1988) studied users' evaluation criteria for information. Using interviews, the authors gathered information about users' situations related to their serious life and health problems. The users also described their information needs during these events, plans for information seeking, data sources used, and information expectations. The researchers were able to identify 36 criteria for acceptance or rejection of information, sources, and information-seeking strategies. These criteria included: serendipity, coverage, requirement/need, logical deduction, ease of access, social pressure, uncertainty, trust or respect, time consideration, and others. It is quite clear that the Cranfield experimental approach to IR evaluation could not produce these results.

As a result of this change of viewpoints, new evaluation methods are being proposed (Hersh 1996). These methods are more oriented towards the comprehensive evaluation of the whole IR process and its components, as compared to the previously used predominantly search engine oriented evaluation methods. They include, for example, investigation of the user's information needs (Westbrook 1993; Wilson 1994) and assessment of user ability

to find and apply specific information (Egan, Remde et al. 1989; Hersh, Elliot et al. 1994; Wildemuth, de Bliek et al. 1995; Hersh, Pentecost et al. 1996). Other approaches are the observation and monitoring of user interaction with a system (Shneiderman 1992; Borgman, Hirsh et al. 1996) and "think aloud" protocol analysis (Ericsson and Simon 1993; Kushniruk, Kaufman et al. 1996; Kushniruk and Patel 1998). There is also an example of outcome-oriented evaluation that addresses user satisfaction and system impact in different organizational settings (Lancaster and Warner 1993), including health care (Hersh 1996).

A Concept of "Relevance"

Importance of the concept of "relevance"

The concept of "relevance" has been the basis for research and evaluation in IR since the first experiments in the field. Saracevic (Saracevic 1975) argues that the concept of "relevance" is the main reason for the emergence of information science as an independent area of research. Park (Park 1994) concurs with this view and describes "relevance" as "a fundamental concept of information science" and "a key problem in IR research." Mizzaro (Mizzaro 1997) also characterizes "relevance" as "a fundamental, though not completely understood, concept for documentation, information science, and information retrieval." This view is supported by other researchers (Schamber, Eisenberg et al. 1990; Froehlich 1994; Schamber 1994).

There is a need to better understand what "relevance" means as a concept, what it means to information systems users, and how users search for relevant data and information sources. All these would help to improve information system design, development and evaluation. For example, Schamber (Schamber 1994) identifies relevance as the central concept of information retrieval, and the fundamental measure of its effectiveness. She argues that relevance influences not only design and evaluation of IR systems, but also our understanding of how users seek, analyze, and apply information.

Reviews

Many papers discussing different views on the concept of "relevance" and its application for evaluation of IR in general, and search engines in particular, have been published since the 1950's. The number of publications referenced in reviews of the research of this concept are in the hundreds (Mizzaro 1997). Some of the most important reviews have been written by Saracevic (Saracevic 1975), and Schamber (Schamber 1994). A special issue of the JASIS (Journal of the American Society of Information Science) discussing the concept of "relevance" was published in 1994 (Froehlich 1994).

Mizzaro (Mizzaro 1997) presents the history of the concept of "relevance" through an exhaustive review of the literature. His task was very difficult as there are many papers written about this concept, and authors often disagree about the meaning of "relevance" and its application. Mizzaro presents a framework for a common ground of discussion and illustrates the history of "relevance" by presenting papers in a chronological order. He identifies about three periods: before 1958, between 1959 and 1976, and from 1977 until the present time. Within each period, papers are analyzed according to seven points of view: methodological foundations, different kinds of relevance, beyond-topical criteria adopted by users, modes for expression of the relevance judgement, the dynamic nature of relevance, types of document representation, and agreement between judges.

Different views

In the Cranfield research paradigm, relevance means "on the topic," "on the subject," "aboutness" (Park 1994). Some researchers (Schamber, Eisenberg et al. 1990; Barry 1994) refer to this view as "topical relevance," meaning that to be judged "relevant," the topic of the document has to match the topic of the query. This approach implies a fixed and unchanging relationship between a query and a document. This view of relevance has been intensively criticized since the beginning of IR experiments. In the Cranfield research paradigm, experiments are mostly conducted in the artificial conditions of experimental settings, and real users do not participate in the experiments. "Topical relevance" does not take into account circumstances of a search, including the individual's information need. It does not consider a user's individual characteristics, as well as the dynamic nature of a user's state of knowledge and information needs during a search. Real users also lack clarity about the notion of "topical relevance" (Park 1994; Schamber 1994). Many researchers point to the necessity of user-centered relevance in IR research. For example, Park (Park 1994) emphasizes "the need to develop the concept of user-based relevance for the benefit of users and for the meaningful development of future research in information retrieval." She argues that the nature of relevance has to be studied from the perspective of real users.

Research of relevance assessments by humans stems from two major experimental studies during the 1960's conducted by Cuadra and Katter (Cuadra and Katter 1967; Cuadra and Katter 1967), and Rees and Schultz (Rees and Schultz 1967). These studies resulted in the identification and testing of factors responsible for the most variation in relevance evaluations. Saracevic (Saracevic 1970) summarizes the major factors affecting relevance judgements: judges' subject expertise at the various stages of research; judges' academic and professional training; a document's intended use; and stylistic characteristics of documents. Based on a review of the literature, Schamber (Schamber 1994) presents a table with 80 factors that can affect relevance judgements. These factors are grouped under broad categories of judges, requests, documents, information system, judgement conditions, and choice of scale.

Different aspects of user-centered relevance have been investigated. A concept of "utility" was introduced quite early in IR research (Cooper 1973). This concept includes quality, novelty, importance, and credibility of information from the point of view of a user, as opposed to just topical relatedness. Wilson (Wilson 1973) discusses a "situational relevance" that implies multiple concepts and involves the relation between information and a particular individual's view and situation. Swanson (Swanson 1986) defines "subjective relevance" as the mental experience of an individual person who has an information need. As another important result of the discussion of user-related relevance, the

difference between "relevance" and "pertinence" has been pointed out (Foskett 1972; Kemp 1974; Lancaster and Warner 1993). "Pertinence" requires that, to be judged "relevant," a particular document has a relationship to a user's specific information problem, and that it changes a user's state of knowledge about this problem (Howard 1994).

Park (Park 1993) presents a comprehensive, empirical view of users' criteria in accepting or rejecting a bibliographic citation. Her findings show that psychological, user-based relevance involves an individual's interpretations and complex mental processes that go beyond simple topical relevance in evaluating citations retrieved by real information systems. In their relevance judgements, users usually take into account many different variables. Furthermore, as users encounter relevant citations, their thinking and approaches to their information problems change dynamically through refocusing. Park's results (Park 1993) support the interpretation of "psychological relevance" in IR presented by Harter (Harter 1992). The concept of "psychological relevance" aims toward bringing together different aspects of a user-based relevance. Harter views a retrieved citation as a psychological stimulus. The citation is considered a relevant one if it causes cognitive changes in the user.

Harter also identifies some basic concepts and principles of psychological relevance. A user has assumptions about the world. These assumptions form a

psychological construct. This construct is a dynamic structure that can change when new assumptions are created. Harter views a user's information need as a perceived "emptiness" in this psychological construct. Some new assumptions are made and tested based on results of the IR process. According to Harter, reading a new citation can cause a user to create new assumptions, strengthen, weaken, or discard old ones. All these result in a new psychological construct, where the previous "emptiness" is compensated for by new assumptions about the world. This view is very similar to the one based on the role of an "anomalous state of knowledge" (ASK) in information retrieval (Belkin, Oddy et al. 1982; Belkin 1984).

Some attempts to summarize different views of "relevance" and clarify the usage of relevance-related terminology have been made. Schamber (Schamber 1994) presents three views of relevance: a system view, an information view, and a situation view. The system view is objective and refers to a match between a query and a document. The information view is subjective and refers to the relation between a request for information and a document. The situation view is subjective and refers to the relationship between a user's information need and a document.

Schamber (Schamber 1994) shows that there is an overlap in the terminology used in each of these views. This overlap is a result of "conceptual grey areas" in

our understanding of the concept of "relevance." One of these areas is the role of topicality in relevance judgements. Even researchers that argue for the subjective character of relevance agree that topicality is a necessary, but not sufficient, condition in the relevance judgements of real users (Park 1994). Therefore, there is a connection between objective and subjective views of relevance. While Froehlich argues that "the prototypical core for relevance judgements or the nuclear sense of relevance is topicality" (Froehlich 1994, p.129), Schamber (Schamber, Eisenberg et al. 1990) proposes that the use of topicality in evaluation is a logical step for information systems development. The conclusion made 25 years ago by Saracevic regarding different views of relevance is still valid today:

> "Different views of relevance are not independent of each other. It seems that there exists an interlocking, interplaying cycle of the various systems of relevance (i.e., various systems of measures)... There is *no*, and there cannot be any *one specific*, view of relevance in communication... Many practical problems in information systems and many cases of user dissatisfaction can now be explained as due to the existence of various systems of relevance." (Saracevic 1975, pp.338-339)

Problems

Despite the numerous publications that have discussed and used the concept of "relevance" since the beginning of information science as a distinct discipline, there exists "little agreement as to the exact nature of relevance and even less that it could be operationalized in systems or for the evaluation of systems" (Froehlich 1994, p. 124). Schamber, Eisenberg, and Nilan argue that "an enormous body of information science literature is based on work that uses

relevance, without thoroughly understanding what it means" (Schamber, Eisenberg et al. 1990, p.756). Park supports their view: "The idea of relevance has played a major role in the evaluation of information retrieval, but without a consensus on the meaning of this concept" (Park 1994, p.135). Schamber (Schamber 1994) thinks that the majority of problems are related to the use of different views of relevance, as well as "loose and inconsistent" terminology by different researches. Froehlich (Froehlich 1994) has identified the main factors contributing to "relevance"-related problems in information science:

1. The inability to define relevance
2. The inadequacy of topicality as the basis of relevance judgements
3. The diversity of non-topical, user-centered criteria that affect relevance judgements
4. The dynamic and fluid character of information seeking behavior
5. The need for appropriate methodologies
6. The need for more complex, robust models for system design and evaluation

Some authors believe that evaluation has to involve real users making relevance judgements (Park 1994; Blair 1996; Ellis 1996; Harter 1996; Hersh, Pentecost et al. 1996). Others argue that any person who is knowledgeable about the subject of a search can make relevance judgements (Salton and McGill 1983; Cleverdon 1991; Salton 1992). This argument addresses the issue of "relevance judgement." The authors supporting the involvement of real users in evaluation consider a

relevance judgement to be a representation of the value of the document for a particular user in a particular situation (time, place, context, etc). The contra-argument is based on the view of a relevance judgement as a decision about whether or not a retrieved document answers a query.

A variety of methods for capturing relevance decisions have been proposed (Salton and McGill 1983; Meadow 1992; Lancaster and Warner 1993; Schamber 1994; Ellis 1996; Harter 1996; Harter and Hert 1997). Some of them use binary ("yes" or "no") or category scales. However, the use of these scales for relevance decisions often implies a direct topical relation between a query and a document based on a fixed and unchanging relationship. The majority of researchers agree that the traditional experimental approach (objectivist approach) is not an adequate methodology for understanding the nature of psychological relevance. The large number of variables that can effect relevance judgements (Schamber 1994) emphasize the necessity of a new methodology for evaluating relevance. As an alternative to the traditional experimental approach, different methods of qualitative research are used for studying user-based relevance. In fact, Park (Park 1994) considers the qualitative methodology as one of the best means for discovering users' criteria for searching, analyzing, and using information sources.

The main advantage of the qualitative research approach (subjectivist approach (Friedman and Wyatt 1997)) is that it takes into account the dynamics of an individual's cognitive state and situational factors. While qualitative research is complex, it can, nevertheless, produce data that is systematic and measurable. All these factors explain the growing popularity of the qualitative research in evaluating relevance in IR (Schamber, Eisenberg et al. 1990; Fidel 1993; Park 1993; Barry 1994; Bruce 1994; Froehlich 1994; Hersh, Elliot et al. 1994; Park 1994; Schamber 1994; Sutton 1994; Hersh, Pentecost et al. 1996; Fidel and Crandall 1997).

Summary

Despite steady progress, the field of IR evaluation continues to have serious problems:

1. The TREC experiments showed that there is a need to be able to use different search engine functions for different types of user information requests (Smeaton 1993). A solution to this problem will require a different evaluation methodology, as none of the existing measures can provide a basis for comparison.
2. As there are many different approaches to IR evaluation, there is uncertainty about the best choice of evaluation methodologies for a specific study. There

is a need for a comprehensive evaluation framework that will serve as a tool for planning IR evaluation, guiding a choice of the most suitable approaches, as well as serving the analysis and comparison of results (Tague-Sutcliffe 1996).

3. Relevance-based measures of recall and precision are not a reliable basis for a decision about superiority of one search engine over another and should be complemented or replaced (Harter 1996). New approaches have to be able to capture in greater detail not only a user's judgement of whether a retrieved document is "relevant" or "not relevant," but allow for the assessment of the complex match between a user's information need (as expressed in a query) and the information in retrieved documents.

In information retrieval, researchers have to be very precise regarding what is evaluated and how the evaluation is performed. "What" relates to defining the boundaries of IR and the characteristics being measured/evaluated. "How" is related to using (1) specific measure(s) that can answer evaluation questions, and (2) appropriate evaluation methods that do exactly what we want them to do.

CHAPTER 4: THE SYSTEMS ANALYTIC APPROACH

The last 100 years have seen the concepts of systems gaining the increasing attention of researchers in different areas of study. The literature discussing theoretical and practical aspects of systems thinking and the systems approach is quite extensive (von Bertalanffy 1976; Blauberg and Sadovsky 1977; Bowler 1981; Checkland 1981; Group 1981; Rapoport 1986; Mesarovic and Takahara 1989; Salvendy 1997). Systems concepts are used in such diverse areas of study as cybernetics, systems engineering, general system theory, operations research, systems analysis, computer systems, and also in fields like geography, planning, accountancy, social work, and psychology. One of the reasons for this growing popularity is that the systems approach uses methods of analysis and design that are applicable to addressing the complex behaviours of natural as well as man-made systems, and their technological and social aspects. Researchers, engineers and policy makers use the systems approach in order to predict and control the behaviour of these complex and diverse systems.

At the same time, our knowledge about systems, systems analysis, control over systems behaviour, or systems design, is still quite limited. There are many examples of unsuccessful attempts to control the behaviour of systems, either natural or human-made. Moreover, researchers often argue about definitions of systems concepts. As neither space, nor the topic of this thesis permits a detailed

discussion of systems concepts, Appendix A presents a short glossary adapted from the work done by the Open Systems Group (The Systems Group of the Technology Faculty at the Open University in London) (Group 1981). I have used their terminology in this thesis.

Czaja argues that according to the systems concept "components or elements of a system are only meaningful in terms of the whole system" (Czaja 1997, p.17). Elements of a system have to be considered from the point of view of their interaction with other elements of the system. The alternative to the systems approach is a reductionist approach, which considers system components or elements as independent, in isolation from each other, and often as stable artefacts. The reductionist approach is often used in the design of computerized systems. Its main focus is on technical components and does not consider behavioral and environmental factors. As one of the results of the reductionist approach, management often gives little or no consideration to human factors during an implementation process.

The systems analytic approach considers the structure, functions, and interactions of the components of a system relative to system goals when evaluating particular phenomena. This approach infers that "performance must be evaluated in terms of the context of the human-machine system; equipment, environment, operating procedures, and goals" (Czaja 1997, p.18). The systems

analytic approach provides a unifying framework for analysis, design, and evaluation of technical and human components of a system and their interaction.

As our understanding of IR grows, it is accompanied by understanding about the enormous complexity of IR and its evaluation. A systems analytic approach to the evaluation of IR can provide a basis for dealing with this complexity. Some examples of attempts to introduce a systems analytic approach in IR are:

- Creating models of IR;
- Identifying boundaries of a system under evaluation;
- Identifying variables under investigation; and
- Creating evaluation frameworks.

Models of Information Retrieval

Many different models of information retrieval (IR) have been proposed. These models show different aspects of IR and reflect the historical dynamics of IR understanding as well as the research interests of their authors. Every model has its strengths and weaknesses. In this chapter, different models of IR are presented, and their strengths and weaknesses are identified and discussed.

General Characteristics of a Modelling Process

Models are useful tools for representing reality. They capture commonalities of similar entities and at the same time can represent specifics as well. People use

models to improve understanding of reality, for analysis of processes and their aspects, for predictions, and other purposes. Models generally include maps, three-dimensional models of cars, buildings, body organs, cells; mathematical models and functional interactive computer models. Tague, Salminen and McClellan characterize the systems modelling process:

> "The purpose of a formal system model is to describe the common features of a set of systems which have been developed for similar problems. The model will explain the structure and processes of these systems, and clarify their general, as opposed to specific, characteristics. The components of a model must include the kinds of entities, relationships, and transformations or operations which form a part of the systems which it is intended to describe. A complete model will contain representation of all components in any system of the kind referenced by the model" (Tague, Salminen et al. 1991, p. 14).

There are several models that are good representatives of the existing modelling approaches. Some of them model what the researchers call an "IR system" (Salton and McGill 1983; Tague, Salminen et al. 1991), others look either into an IR process (Croft 1993), a user (Marchionini 1992), or an information flow in the world (Meadow 1992).

A Search Engine Model

Salton and McGill (Salton and McGill 1983) present a focused view of the technical IR system. Their model represents a functional approach to IR, and specifically the matching process of database records and user queries. Salton's early work (Salton and Lesk 1968) provides the essence of this model: "Every

information retrieval system can be described as consisting of a set of information items (DOCS), a set of requests (REQS), and some mechanism (SIMILAR), for determining which, if any, of the information items meets the requirements of the requests" (cited in (Salton and McGill 1983, p. 10)).

The set of information items is not compared directly with the set of requests. A comparison is possible because both sets are transformed into a special representation form using the same indexing language (LANG). Indexing languages try to capture text semantics[3]. Representing information items using the indexing language is called indexing. A representation of user requests using the same language constitutes a query negotiation process. After this, the representation of the information items is compared to the representation of a given request through the similarity measuring process (SIMILAR). This process can be considered a retrieval function, since it identifies the specific information items that are to be retrieved and presented to a user. Unfortunately, Salton and McGill do not represent user involvement in the information retrieval process, computer-user interface, or environmental influence. Both authors have their background in computer science and information systems; for example, Gerard Salton was interested primarily in the area of improving technical IR systems (search engines), and specifically in indexing and retrieval algorithms.

[3] In our work we discuss different approaches to capturing text semantics during indexing: natural language processing, knowledge representation, and statistical (Kagolovsky, Miller et al. 1997).

A Formal Model of a Search Engine

Tague, Salminen and McClellan (Tague, Salminen et al. 1991) have created a formal model of an IR system. This model addresses the physical and logical system used to provide information in response to user queries. The authors discuss earlier Boolean, vector space, relational, and semantic models and argue that "IR models differ in the components which they incorporate and in the discriminations which they make. Most are based on a mathematical formalism which permits a description of the relationships and operations of the model. Often, though this formalism cannot describe important features of actual systems" (Tague, Salminen et al. 1991, p. 15).

They then describe a model that they developed using two formalisms: production grammars and hypergraphs. They argue that their proposed model provides a more complete basis for the description of modern full-text IR systems by incorporating representations of indexing, ranking, and navigating. However, the proposed model does not deal with non-text systems. It includes neither the cognitive aspects of the process, such as query negotiation or output evaluation, nor computer-user interface or environment.

An Information Retrieval Process Model

Croft (Croft 1993) provides a model of basic information retrieval that consists of three processes: a representation of a user's information problem or need in a form of a query, a representation of text documents using indexing, and a comparison of these representations. This scheme emphasizes the retrieval task. The results are represented in a set of retrieved documents. Comparison is most effective when a query is composed using the same representation method that was used for indexing documents.

Croft's scheme is very close to that of Salton and McGill's model (Salton and McGill 1983). At the same time, Croft emphasizes that representation of an information need is a dynamic process involving user feedback after the evaluation of retrieved documents. This process can result in changes to an information problem and its representation, which often causes the query to be reformulated. Although Croft mentions the interaction between a user and an intermediary, user involvement in the information retrieval process is not explicitly specified, and characteristics of a computer-user interface and an IR environment are not discussed.

A Model of an Information Flow in the World

The model of the IR process presented by Meadow (Meadow 1992) is the most comprehensive one among those discussed here. The author begins by presenting a general view of the setting of information retrieval systems and their relationships to other IR components such as databases, users, and the world. In this model, information about the world comes from communities of current or potential users, and is employed by other users to affect the world.

> "[IR] is a process that depends on separate actions of two groups. *Users* need information and use it to affect future events in the world. *Database producers* study what kind of information is available and what is needed. Then they collect and use it to create the records of a database. The user's information need may be conveyed to an *intermediary*, a person whose profession it is to assist users. The provision of information by the database producer and its search and retrieval by the user are asynchronous. Both will fare better if they have some understanding of each other: the user, of what producers' policies are, and the producers, of users' general needs and characteristics" (Meadow 1992, p.4)

The base model represents the cycle of information flow between information producers, technical IR systems, users, and the world. It encompasses world events that trigger written observations. These observations are gathered by database producers and put into a database form that is connected to a technical IR system. Meadow also specifies components (programs) of a search engine that perform functions of file update, and data and query management. Users, either directly or through an intermediary, interact with a search engine by means of queries. A technical IR system retrieves information items from a database and presents them to a user. Users' work with retrieved items may

produce new observations for the world. Database producers improve functioning of a search engine through the gathering of information from database users.

After presenting a base model of information flow in the world, Meadow discusses in detail the steps involved in communication between a user and a search engine, the functions of a database producer, and major program components of a search engine. But although the coverage of IR components' structure and the relationships between them is very detailed, he considers neither a computer-user interface, nor characteristics of an environment in which IR happens. Moreover, he does not discuss the possibility of using this detailed model for improving evaluation of information retrieval.

A Model of the User

Marchionini's (Marchionini 1992) model addresses the information-seeking behavior of a user. The goal of this model is to provide a basis for creating and evaluating computer-user interfaces to support a user's information seeking. Marchionini considers information seeking as a form of problem solving. The main functions of the information seeking process are problem definition, source selection, problem articulation, examination of results, and information

extraction. Usually, we consider the information-seeking process as iterative with functions following each other in the above order. The end of one information seeking cycle may constitute the beginning of another one. In reality, though, the sequence of the functions performed during the information-seeking process can vary. For example, a user can change an information need and re-define a problem at any stage. This can occur during query formulation (problem articulation) as the result of interaction with an IR system or an intermediary (e.g., a librarian), as well as during an analysis and synthesis of retrieved information (examination of results and information extraction). In Marchionini's model, these functions are represented in a non-linear way. The sequence of the functions depends on the dynamics of the specific user's information-seeking behaviour. However, the author does not address relationships between a user and other components of IR.

Boundaries

In their paper "On the Evaluation of IR Systems," Robertson and Hancock-Beaulieu (Robertson and Hancock-Beaulieu 1992) discuss the increasing complexity in the evaluation of IR systems. They identify the following contributing factors to this increasing complexity: the relevance revolution, the cognitive revolution, and the interactive revolution. The authors argue that

identifying the boundaries of a system under evaluation is one of the central problems that has to be explicitly addressed in order to improve IR evaluation. The authors explain that at the beginning of IR experiments in the 1950's, the system was primarily limited to the retrieval mechanism and such human activities as indexing and searching; input was the request to this mechanism, and output was the retrieved items. The system was usually considered a "black box" with inputs and outputs. Experiments that follow the Cranfield research model use this view of an IR system.

During the last 40 years, this view has been revised considerably. Robertson and Hancock-Beaulieu argue that our research of relevance, cognitive processes involved in information seeking, and user-computer interaction, have shown clearly that boundaries of an IR system under evaluation have to be reconsidered. The boundaries depend on the goals of evaluation, the IR components under evaluation, environment, and other factors. The boundaries of the system under evaluation have to be explicitly identified in every IR experiment.

Variables

In any kind of controlled experimentation, researchers try to identify independent and dependent variables, and evaluate relationships between them. Usually, the number of variables is limited so that meaningful conclusions may be drawn from the experiments. After experimenters choose a proper model of IR and boundaries of a system under evaluation are defined, the next logical step is a process of identifying different variables for investigation.

Fidel and Soergel (Fidel and Soergel 1983) review the literature and collect different factors that have been found to affect online information retrieval. These factors (independent variables) are organized in a conceptual framework. This framework is justified by a need to conduct controlled evaluation of IR and its components. The authors argue that their framework can be used "for integrating the results of previous studies and for guiding future investigators in their choice of research problems and variables so that studies might more easily form a cumulative body of knowledge" (Fidel and Soergel 1983, p.163). In this framework, variables are divided into two groups: independent and dependent variables. Fidel and Soergel also discuss how different variables in the IR process can be viewed, and present a typology of roles variables can play in a study. I concur with the authors that in IR evaluation we are interested in identifying the effects of independent variables on the elements of the IR process and its outcome, and that "[m]uch would be gained if researchers would clearly state

what the dependent variables in a study are, what variables might affect the dependent variables, and what role each of them play in the study" (Fidel and Soergel 1983, p.164).

Fidel and Soergel consider the following eight elements as parts of IR: the user, the request, the database, the search system, the searcher, the search process, and the search outcome. After a detailed discussion of their framework, a list of variables, excluding search outcome, is presented for each of these elements. The authors argue that in analyzing a specific IR process, the values of variables that characterize each of these eight elements have to be known. They also advocate a use of specific combinations of elements, for example, "the match between cost restrictions imposed by the setting and the cost of searching a database, or the familiarity of a given searcher with a given database" (Fidel and Soergel 1983, p.164). Mechanisms by which combination variables can be generated are presented.

These are some of the main problems associated with this attempt to organize factors affecting IR into a conceptual framework:

1. Although the list of variables is quite large, the authors realize that it is not exhaustive. They consider this list as a generator of variables. At the same time, the authors admit that an exhaustive list of variables is almost impossible to create since IR involves human and organizational elements.

The number of factors that can influence human and organizational behavior is practically unlimited and many of them can affect the IR process.

2. The authors also found that individual variables are not very useful for characterizing IR. As an alternative, the authors advocate a more complex approach to variables: "With a few exceptions such as cost-consciousness, individual variables taken alone seem to have little influence; exploration of combination variables and of even larger and more complex patterns, hold more promise for understanding the search process and its outcome" (Fidel and Soergel 1983, p.169). However, I think that a more complex approach undertaken without using a structured model of IR will further complicate the matter of IR evaluation.
3. Although this conceptual framework tries to establish an order necessary for a controlled evaluation of IR, I found it quite complicated. It is probably related to the level of complexity associated with IR and its evaluation. However, merely knowing variables of the IR process does not necessarily result in an improvement of IR evaluation. Without using a structured model of IR, any list of variables can be difficult to use for analysis, design, and evaluation of IR experiments.

Evaluation frameworks

As IR is a very complex process, different methods of evaluation have been proposed that evaluate different aspects of information retrieval. To organize these methods, as well as to assist in using them in evaluation, some evaluation frameworks have been proposed. For example, Lancaster and Warner (Lancaster and Warner 1993) propose a classification that evaluates information services on the basis of cost, time, and quality. The authors identify three possible levels of evaluation: effectiveness, cost effectiveness, and the cost-benefit relationship.

The first level, the evaluation of effectiveness, is related to evaluation of the search engine, and of the user interacting with it. Three groups of criteria are considered: cost, time, and quality considerations. Criteria of cost and time are relatively straightforward and easy to compare. Two types of cost are evaluated: direct cost per search, document, or subscription, as well as indirect cost considerations related to efforts involved in learning to use the system, actual use, and retrieving information sources. The following time criteria can be of interest for evaluation: how long the user has to wait to use the system, how long it takes from the submission of a request to the retrieval of references, and also to the retrieval of documents. At the same time, the criteria of quality are much more subjective and can vary significantly depending on the system evaluated and the user's information need. Lancaster and Warner include in this category such measures as the coverage of a database, completeness (recall) and the

relevance of output (precision), the novelty of retrieved sources, and their completeness and accuracy.

The second level relates measures of effectiveness to measures of cost. For example, it evaluates the unit cost of different aspects of information retrieval results: a relevant citation retrieved, a previously unknown relevant citation retrieved, or a relevant document retrieved. Usually, users are not aware of costs. However, people responsible for the financial status of organizations are interested in making sure that money is well spent.

The third level of evaluation relates the cost of providing information services to the benefits of having the services available. At this level, researchers are interested in assessment of how money invested in information technology results in the improvement of processes, for example, better decision-making or outcomes. This type of evaluation is very important, because it addresses changes in organizations and their functionality. However, this kind of research is often difficult to conduct, as it requires large-scale data collection and interventions in operational settings.

Conclusion

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 have been identified by several researchers (Fidel and Soergel 1983; Salton and McGill 1983; Robertson and Hancock-Beaulieu 1992; Tague-Sutcliffe 1992; Croft 1993; Tague-Sutcliffe 1996). 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 a better understanding of the IR process. Some researchers do not discriminate between "retrieval system evaluation," "retrieval system performance", evaluation of "information retrieval procedures," and "retrieval evaluation" (Salton 1992). This results in terminological differences and the inability to compare 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.

Models of information retrieval have to accommodate a naturalistic setting and focus of users' information needs (Froehlich 1994). To be comprehensive, IR

models have to include users' mental models (Belkin 1984; Sutton 1994), information seeking behavior (Howard 1994), the iterative character of information seeking (Hersh 1994; Hersh, Pentecost et al. 1996), a cognitive model of a user's interaction with an information retrieval system (Bruce 1994), and other user characteristics.

The systems analytic approach allows us to combine the lessons learned from these different ways of looking at IR: modeling, identifying boundaries, working with variables, and classifications of evaluation methods. We need a structured model that presents components of IR and interactions between them, as well as a robust method for identifying boundaries of a system under evaluation, and for investigating variables of the IR process. This systems analytic approach has to permit the analysis of existing methods of evaluation, as well as to propose new evaluation methodologies.

CHAPTER 5: CONCEPTUAL IR FRAMEWORK

Identification of IR components, their boundaries, and relationships

A generalized model of information retrieval may be stated as follows. Humans initiate the IR process. A user operates in an **environment**, which is characterized by such variables as **setting** (laboratory or operational), **organizational policy**, or **type of research**. The user encounters a problem that she has to solve (make a decision, take an action) in an environment. To be able to do this, the user accesses and evaluates her **state of knowledge (SK)** about the task. This knowledge is expressed as a cognitive structure in a form of a propositional matrix (Kintsch 1998). If a user's SK is not sufficient for decision-making, the user has an **anomalous state of knowledge (ASK)** (Belkin, Oddy et al. 1982; Belkin 1984). This state of knowledge can also be called an implicit information need. Based on the ASK, the user can make a statement about her information need explicitly using a natural language. However, it has been shown that users usually do not sufficiently translate their implicit information needs (ASK) into natural language (Schamber 1994). As a result, a formulation of a user's information need can vary from a very vague to a quite precise one. It depends on the clarity of a user's cognitive model and language abilities. Some environmental variables can play role too, for example, location, time, and cost constraints.

The next step is to make a decision about a strategy to resolve the ASK. Some authors state that a user does not look for definite answers, but rather for a "treatment" of an ASK (Tague and Schultz 1989). This decision is based on a user's previous knowledge about the subject, formulation of information need(s), availability of information sources, a user's familiarity with these sources, and other factors. Some of the possibilities include looking for a person who has the needed information, or finding this information in a printed form. A search for printed resources can lead to library and/or electronic data sources.

In some situations, users will perform searches and evaluate results by themselves; in others, information needs are explained to a searcher who performs a search. If a user contacts an intermediary, it depends on the user's ability to explain her information request, as well as on the intermediary's ability to conduct an interview, the intermediary's knowledge of information resources, and other factors. The user expresses an **information need** that results in a **query formulation** and the building of a **search strategy**. All these processes depend on such **characteristics** as a **state of knowledge** in a specific field and **searching skills** (level of training and experience).

The human being interacts with the technical IR system through the **interface**. The interface is any medium that transforms information queries into system specific commands and presents the retrieved set of documents. This bi-

directional function explains the relation between the processes of the user's query formulation and the querying mechanisms of the technical IR system. These mechanisms work in connection with other functions of the technical IR system. Human beings have to be able to transform their information needs properly into a **search statement** (the verbalized statement of information needs) and build a **query** through the use of the interface. The **query** 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.

A **technical IR system** (**search engine**) is a computer-based information system used by humans 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. They interact to retrieve documents from a **document set**. Thus, retrieval can be generally understood as the process of comparison of the query with the indexed documents in the document set.

The **document set** consists of documents that are chosen based on a speciality (e.g., pediatrics), subject coverage (e.g., Canada), type of resources (e.g., journal articles, news, etc.), and other criteria. The document sets are usually created by

a vendor or an experimentator. New documents are added to the set periodically. If a document set is used in IR experiments, experts usually define the number of documents in the set, their characteristics, structure, and coverage. Experts also usually identify sub-sets of documents that are judged "relevant" to a query, i.e., a "relevance base."

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 on a screen or as a printed copy. This set of documents can be read and evaluated by a user who extracts information from the documents and attempts to incorporate it into her cognitive model. Based on this cognitive model, the user can either decide to stop her search or to adjust her information needs. In this last case the cycle outlined above can be continued. However, it has to be mentioned that 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.

Proposed models of IR

The proposed structured model (Fig. 1) is a graphical representation of all the major components of the IR process and relationships between them. The relationships between components follow the steps of the IR process. Some of the internal characteristics of a search engine, and of a user, are identified. The decision to identify these particular components was based on availability of space in this figure. There are many more characteristics that can be specified for every component of IR. Other versions of the same model can be created to show different components and/or steps of the IR process in more detail. I have also created a process model of IR(Fig. 2). This model shows steps a user takes during the IR process. It is similar to a flow-block diagram according to the definition by the Open Systems Group (The Systems Group of the Technology Faculty at the Open University in London) (Group 1981).

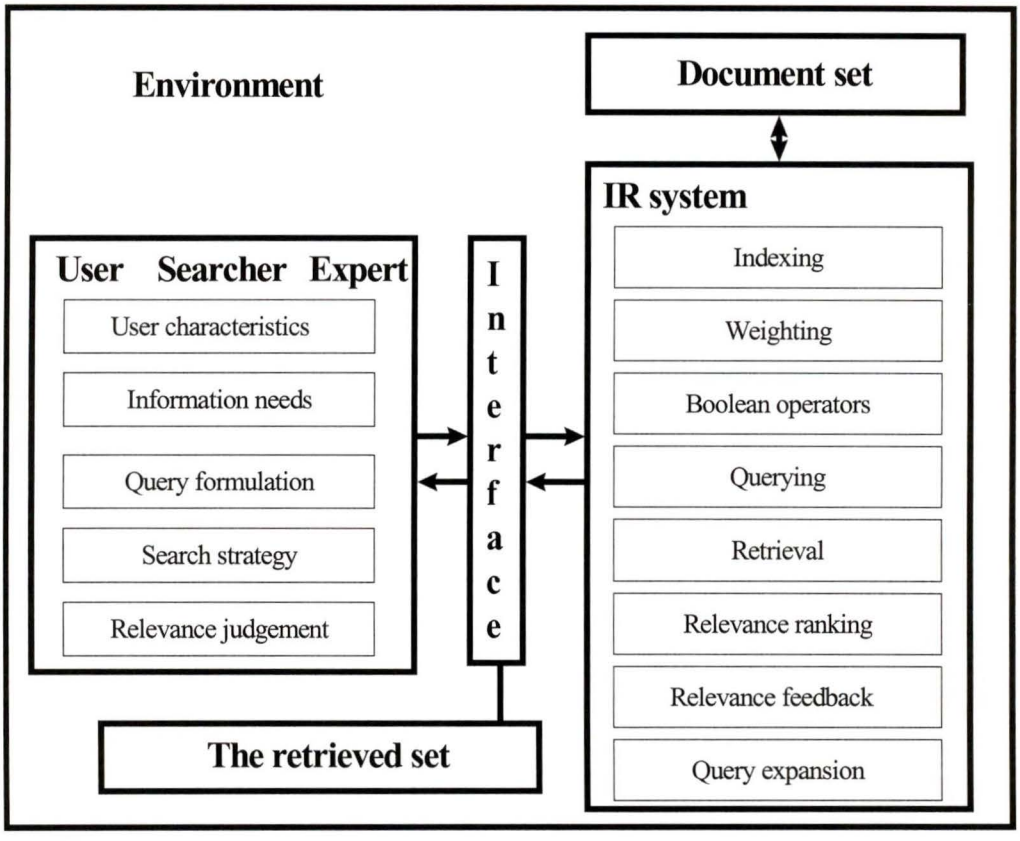


Figure 1. Components of the IR process

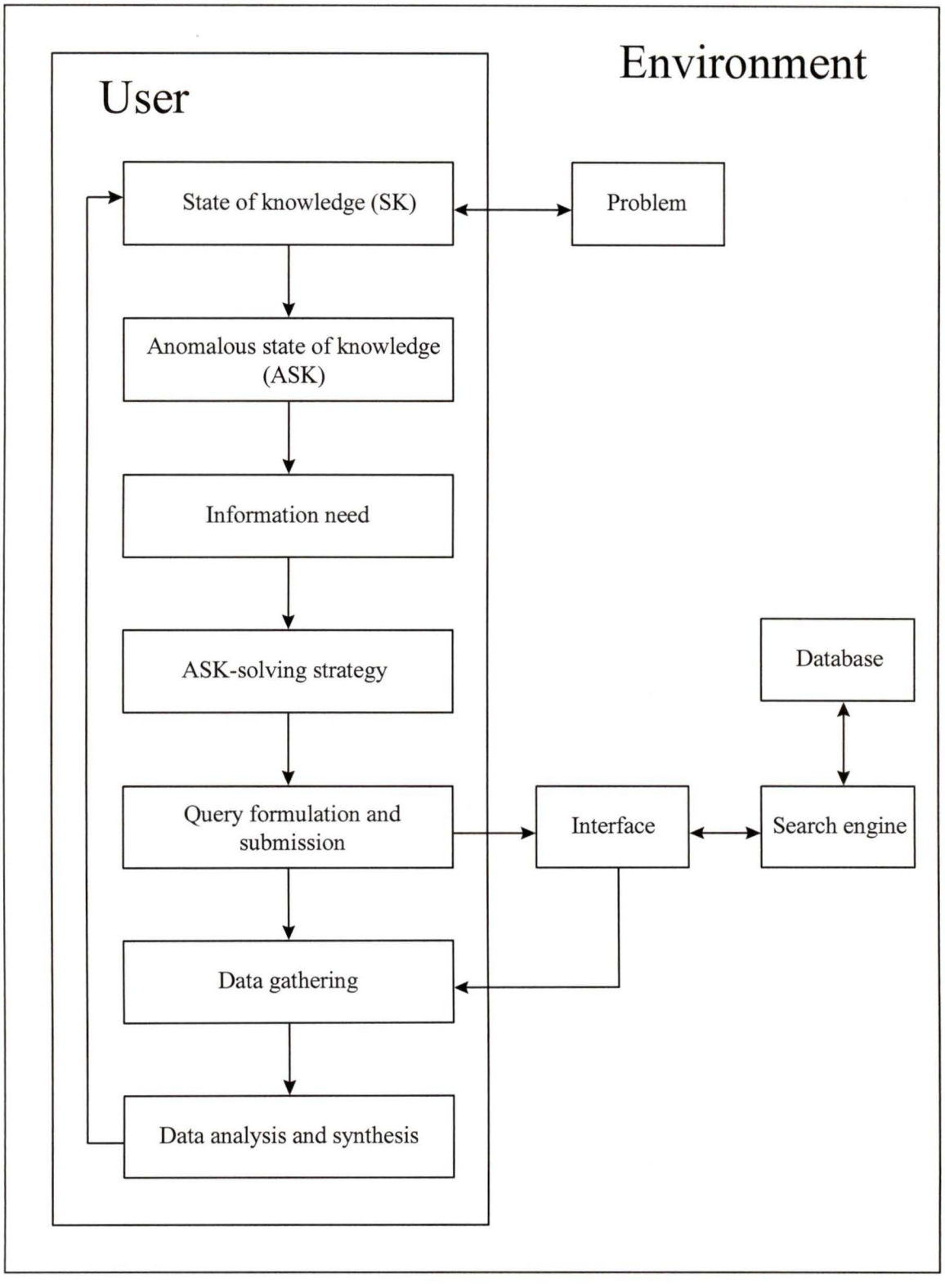


Figure 2. Process model of IR

CHAPTER 6: IMPLICATIONS

A definition of IR

I 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. This definition uses the proposed structured model of IR, states the interactive character of IR, specifies its dynamic nature, and includes components of the IR process and relationships between them. Other important aspects are the explicit identification of an IR goal, the inclusion of a notion of relevance between information sources, and a user's information need. Therefore, my definition identifies information retrieval as a more comprehensive process than just a retrieval of documents using computers.

An introduction of common terminology

Another result of applying a systems analytic approach is the introduction of a common terminology that can provide researchers with grounds for a better understanding of the goals and objectives of work done by other authors. Such terminology can also improve understanding of the results of different experiments and their comparison.

An example of introducing a common terminology is related to understanding and using the term "retrieval." The proposed structured model permits researchers to be more precise in using such terms as "information retrieval," "information retrieval process," "technical information retrieval system" ("search engine"), "retrieval process," and "retrieval algorithm." This precision can be achieved as a result of applying a systems analytic approach that requires identification and definition of concepts used during analysis. The terms "information retrieval," "information retrieval process," and "retrieval process" can be used interchangeably, as they refer to the same dynamic process that includes interaction of all IR components. All these terms have a meaning different from the terms "technical information retrieval system" or "search engine" that refer to one of the IR components. The term "retrieval algorithm" has to be used when discussing functions (characteristics) of search engines, in order to avoid using the ambiguous term "retrieval."

Relevance-based measures of recall and precision

An introduction

The main goal of this discussion is to identify what components of IR are evaluated by relevance-based measures of recall and precision. Recall and precision have been generally accepted as evaluation measures for search engines since the first experiments in IR, when the main focus of a research was on improving indexing and retrieval algorithms (Cleverdon 1970; Salton and McGill 1983; Cleverdon 1991; Meadow 1992; Salton 1992; Harter 1996; Hersh 1996; Harter and Hert 1997). Although relevance-based evaluation methods have been intensively criticized, the critique has been related primarily to a discussion about the meaning of the concept of "relevance" and the implementation of a relevance-based evaluation -- how to understand "relevance" and how to conduct experiments using relevance-based measures. At the same time, I have not been able to find any explicitly stated argument against the notion that recall and precision are evaluation measures of search engine performance.

I am going to demonstrate that relevance-based measures of recall and precision are evaluation measures of the whole IR process, and not just of search engines. I will investigate this problem using the structured model of IR. To begin, I will look at the different views on search engines as a part of the IR process, and the implications of these views for search engine evaluation. After this, the

theoretical basis of relevance-based measures of recall and precision will be explained based on the general principles of experimental design. Next, my critique of this theoretical basis will be presented.

Two views of a search engine

According to the structured model, a search engine is a specific type of computerized system, the primary functions of which are indexing and retrieval of documents. Internally, it can be characterized as a computerized system with specific characteristics: for example, speed, indexing and retrieval algorithms, stop-word lists, and results representation. Externally, it interacts with users, accepting queries and producing results that may or may not be evaluated by users. Therefore, a search engine has to be evaluated both from the internal and external points of view. An internal evaluation is represented by the methods usually used in computer science (Salton and McGill 1983; Meadow 1992; Kantor 1994; Hull 1996; Harter and Hert 1997; Greenberg and Garber 1999). An external evaluation has historically been accomplished using relevance-based measures of recall and precision.

The theoretical basis of recall and precision measures

The theoretical basis for the relevance-based evaluation measures of recall and precision can be explained using the proposed structured model. In a typical Cranfield-like experiment, the measures of recall and precision are used to make a decision about the superiority of one search engine over others based on a comparison of differences in retrieved sets[4]. According to the general principles of experimental design, such comparison can only be possible if all other structural components of IR (i.e., environment, user, interface, document set) are kept constant, and changes are introduced in the algorithms of search engines. If some other variable(s) in addition to search engines are not constant, measures of recall and precision evaluate the whole IR processs, rather than just evaluating search engines.

A critique

Although such components as environment, interface, and document set are easy to keep constant in laboratory settings, the notion of a "constant user" is a problematic one. A Cranfield-type experimental design cannot keep a user component "constant." A user interacts with a search engine through a

[4] For the purpose of this discussion, introducing a change in the settings of a search engine is equal to introducing a new search engine.

computer-user interface by submitting a query and receiving retrieved documents for evaluation. Therefore, an "ideal" user has to produce the same query and always evaluate results the same way. The input part (query) is not difficult to hold constant. In contrast, results evaluation depends on a user's "relevance" judgement. There are different approaches to the treatment of "relevance" as a concept, varying from objective to subjective; but in general, research has shown that relevance is subjective and depends on many different factors (Schamber, Eisenberg et al. 1990; Froehlich 1994; Schamber 1994). This is exactly the reason why, in an experiment with real users who are making relevance judgements regarding retrieved documents, the agreement between users is not very high. Therefore, in such experiments it is impossible to make a comparison between search engines based on measures of recall and precision, as there is at least another component that is not constant in addition to a search engine -- a user. Therefore, measures of recall and precision evaluate the whole IR processs, rather than just evaluating search engines.

Cranfield-like experiments attempt to solve this problem first by not using real users, and second, by establishing a constant "user's" decision about the relevance of every document in a dataset in relationship to a specific query. However, this solution has some serious problems:

1. It has been argued many times that the notion of a "constant" relevance judgement is an artificial one, and that results of such experiments cannot be extended to operational settings. The use of "relevance" as an absolute, unchangeable measure does not correspond to the real world situation (Schamber 1994; Harter 1996). First, a user's relevance judgement is based on her understanding of the information need and level of subject knowledge, and will, therefore, be different for different users. Second, real-life IR experiments demonstrate that a user's relevance judgement can change during a search through interacting with a search engine, viewing retrieved documents, and so forth. Therefore, evaluation of search engines based on objective understanding of "relevance" is not realistic[5].

2. I also would like to address the issue of trying to identify the superiority of one search engine over others based on differences in retrieved sets. In the context of the TREC experiments it was shown that even search engines with poor overall performance come out on top for certain queries (Smeaton 1993; Harter 1996). Therefore, assessment of the differences in search engines' retrieval performance is important. Unfortunately, it cannot be done using existing evaluation methods.

[5] Recall and precision use objective understanding of "relevance" expressed in a pre-defined "recall base" for each query.

I agree that measures of recall and precision can still be useful during the first steps of search engine evaluation. Such measures can take advantage of the established test collections with queries and corresponding recall bases. However, as has been shown above, recall and precision are not evaluation measures of search engines, but rather of the IR process as a whole. Therefore, Cranfield type experiments and measures of recall and precision used in these experiments cannot be used for comparing search engines. The relative evaluation of search engines proposed by Salton (Salton 1992) also does not address the main problem related to "relevance," which is its subjective character. I argue that new approaches to external evaluation of search engines can be created using the structured model of IR.

Critical analysis of the concept of "relevance"

Although the concept of "relevance" is a central one for the field of IR, there is no clear definition, understanding, or consistent use of this concept. In this chapter, I will present a discussion of "relevance" using the modified structured model of IR (Fig. 2). The goal of this discussion is to show that there are different kinds of "relevance" relationships. This is consistent with Mizzaro's (Mizzaro 1997)

opinion that "there are many kinds of relevance, not just one". Every one of them has to be specified and analyzed using appropriate evaluation methods.

Using the modified structured model of IR, the following "relevance" relationships can be identified and considered for evaluation:

1. How is a user's state of knowledge relevant to a problem?
2. How is a formulated information need relevant to a user's state of knowledge?
3. How is a user's information seeking strategy (ASK-solving strategy) relevant to
    a. a user's explicit information need?
    b. a user's implicit information need (ASK)?
    c. a problem?
4. How is a formulated query relevant to a user's
    a. state of knowledge (implicit information need)?
    b. explicit information need?
5. How are retrieved documents relevant to
    a. a submitted query?
    b. a user's explicit information need?
    c. a user's ASK?
    d. a user's problem solving?

This list of "relevance" relationships includes the most typically used ones. For example, it includes the most often considered views of "relevance" as the "subject relatedness" of a record to a query and the "value" or "utility" of a record to a user. In addition, it includes those relationships identified by other researchers (Mizzaro 1997), such as how retrieved documents are relevant to a problem or a user's ASK. This list further adds to these new types of "relevance" relationships and introduces new perspectives on the concept of "relevance" in IR. For example, the following "relevance" relationships have not been identified by IR researchers: how a formulated information need is relevant to a user's state of knowledge; how a user's information seeking strategy is relevant to a user's explicit information need, user's implicit information need (ASK), and a problem. Although further "relevance" relations might exist, I consider the proposed list as likely complete for now.

Although, some authors describe the "subject relatedness" as corresponding to type 5a in my scheme of "relevance" relationships (Salton and McGill 1983; Salton 1992), others identify it as corresponding to type 5b (Meadow 1992). The same is true for the interpretation of the "value" and "utility" aspects of "relevance" that might correspond to types 5c or 5d, depending on the preferences of researchers. In contrast, my proposed classification of "relevance" relationships permits me to identify more precisely the types of relationships under consideration. Therefore, my structured model of IR built using a systems

analytic approach makes the process of "relevance" analysis and evaluation more comprehensive and precise. Moreover, it will be shown that the same model provides a basis for choosing appropriate approaches to evaluating these "relevance" relationships in IR. These approaches will be discussed later in this thesis.

Chapter 7: Proposal for Solution

In this chapter I would like to propose new approaches to IR evaluation. The focus of the proposed methods is the main goal of information retrieval - the satisfaction of users' information needs.

To improve IR I propose to use methods that can evaluate:

1. "Relevance" relationships between steps of the IR process.
2. User participation in the IR process.
3. Evaluation of search engines based on users' information needs satisfaction.

As it was shown in Fig. 2, the IR process can be represented as a sequence of steps. A result of one step serves as an input to another step. "Gaps" in this sequence result in a failure of the IR process to satisfy users' information needs. These "gaps" can be characterized as a disparity between two steps of IR process. Identification of such disparities serves as a first step to further investigations of their source(s), and an improvement of IR.

To find "gaps" in this sequence, we have to be able to evaluate a "relevance" relationship between steps of the IR process. As a "relevance" relationship exists between results of any two steps of the IR process, evaluation of "relevance" gives information about correspondence between these steps. Such evaluation

involves capturing the semantics (meaning) of results of these two steps in some formal representation and comparing them. This evaluation methodology will permit us:

1. to go beyond existing binary evaluations of "relevance" ("relevant" or "not relevant") used in relevance-based measures of recall and precision, and
2. to apply more sound evaluation approaches that are based on methods of cognitive psychology.

When a "gap" in the IR process is identified, other methods can be used to investigate it. If a "gap" is related to users' involvement in IR, methods of cognitive psychology can be used. For example, although all users have to formulate an information need statement, choose a strategy, or evaluate retrieved documents, these tasks can be done differently by different users. Therefore, if a "gap" is related to one of these steps, evaluators have to understand how users perform these tasks. One possible approach is a "think aloud" protocol analysis that is used in cognitive psychology (Ericsson and Simon 1993; Kushniruk, Kaufman et al. 1996; Kushniruk and Patel 1998). The results of this type of analysis would allow the creation of information systems that can support commonalities and accommodate differences between users.

Another possible "gap" in an IR process can exist if the functionality of a search engine cannot satisfy users' information needs. This requires new methods of

evaluating search engines. These methods have to be based on capturing the semantics of users' information needs and comparing them with semantics of retrieved documents. A discussion of proposed methods, as well as approaches to search engine comparison, will be presented in this chapter.

A proposal for evaluating "relevance" relationships

I propose the following steps to evaluate "relevance":

1. Choose a "relevance" relationship for evaluation.
2. Characterize this relationship.
3. Identify an experimental design; for example: identify subjects, how the experiment will be conducted, and other issues.
4. Capture the semantics of each part of the "relevance" relationship; for example: a query, a document, a user's state of knowledge, a problem description, a user's action, and others.
5. Compare the semantics of both parts involved in the "relevance" relationship.
6. Make suggestions.

One example could be an evaluation of a "relevance" relationship between a problem description and a statement of users' information needs. This relationship reflects how well users are able to express their understanding of the

problem by means of a natural language[6]. Evaluating this type of relationship is very important, as it is well known that people usually cannot sufficiently express their needs for data and information by means of a natural language (Schamber 1994). Thus, librarians often complain that users usually have difficulties formulating what they are looking for during a pre-search interview.

The following is a possible experimental design. Users are presented with a description of a problem[7]. They are asked to "think aloud" as they analyze the problem and identify their knowledge about this problem, trying to formulate a statement of their information needs. A problem description and users' information needs statements are analyzed using, for example, propositional analysis (Kintsch 1985; Kintsch 1989; Kintsch 1998). Their captured meanings are compared and a conclusion about this "relevance" relationship is made[8].

The proposed experimental design would also permit analysis of a dynamics of working with cognitive structures that characterizes users' states of knowledge about a problem. This can be done using "think aloud" protocol analysis (Ericsson and Simon 1993; Kushniruk, Kaufman et al. 1996; Kushniruk and Patel 1998). The results of this analysis can help to identify problems and cognitive

---

[6] As we are interested in a user's ability to translate mental constructs into a natural language, we do not have to investigate a user's original level of knowledge about the problem.

[7] Although a problem can be also presented through a story-telling or a movie, a textual form is easier to use for our purposes. However, in a future it could be interesting to investigate different ways of presenting a problem.

blocks, and to plan steps for improving a process of formulating users' information needs statements, for example, users' education.

A proposal for improved evaluation of search engines

As has been shown earlier in this thesis, there is a need to improve evaluation of search engines. Usually methods of computer science are used to evaluate search engines internally -- their characteristics as computerized systems. An external evaluation is intended to provide information about a search engine's ability to satisfy user's information needs. However, currently used relevance-based measures of recall and precision are not sufficient for external evaluation. Reasons for this non-sufficiency were discussed previously in this thesis.

I argue that a comprehensive external evaluation of a search engine has to address the following questions:

1. How are document(s) retrieved by a search engine relevant to users' information needs?
2. How well does a search engine satisfy diverse user information needs?
3. How are search engines different in satisfying users' information needs?

---

[8] Methods of semantics capturing and comparison are discussed later in this chapter.

The **first question** addresses the "relevance" relationship between the retrieved document(s) and the users' information needs. The evaluation of "relevance" relationships was discussed earlier in this chapter. As users usually have problems formulating an information needs statement, choosing a right strategy, and composing a query, it was argued that it can be advantageous if search engines can retrieve documents that cover a variety of topics related to users' information needs (Smeaton 1993). Such functionality might help users to revise their search. This aspect of search engine performance is addressed by the **second question**. The **third issue** is related to a comparison of search engines based on their ability to satisfy users' information needs.

At the same time, users' information needs can exist in three forms (Fig. 2):

1. Cognitive structures of ASK.
2. The same ASK expressed in a natural language (an information needs statement - INS)
3. A query, when a search engine is used.

Therefore, when conducting an external evaluation of search engines, we have to specify a problem addressed (one of the three questions) and the form of users' information needs considered. This results in nine possible approaches to the external evaluation of search engines. For convenience of discussion, and for planning of future experiments, these approaches are represented in Table 1.

	Query	Statement	ASK
Question 1	A	D	G
Question 2	B	E	H
Question 3	C	F	I

Table 1. Possible approaches to the external evaluation of search engines

Methods A, B, and C can be used for an evaluation of search engines when users' information needs are expressed in the form of a query.

Methods D, E, and F focus on evaluating search engines when users' information needs are expressed in the form of an information needs statement (INS).

Methods G, H, and I can be used for evaluating search engines when users' information needs are expressed in the form of a user's ASK.

Methods A, D, and G focus on evaluating "relevance" relationships between retrieved documents and different forms of expressing users' information needs (A – query, D – information needs statement, and G – users' ASK).

Methods B, E, and H can be used for evaluating a search engine's ability to retrieve documents covering different aspects of a search topic when users' information needs are expressed in different forms (B – query, E – information needs statement, and H – users' ASK).

Methods C, F, and I focus on comparing search engines' abilities to satisfy users' information needs expressed in different forms (C - query, F - information needs statement, and I - users' ASK).

To improve the evaluation of search engines, I propose the application of the following **evaluation strategy**:

1. Evaluate search engines and not the whole IR process.
2. Capture semantics (meaning) of users' information needs.
3. Capture semantics (meaning) of retrieved document(s).
4. Compare captured semantics (meaning).

Evaluating search engines and not the whole IR process

As has been discussed in the chapter critiquing relevance-based measures of recall and precision, to be able to evaluate search engines reliably, we would have to keep as many variables constant as possible, including users. In this case, the only component of the IR process that can be changed is a search engine. The effects of these changes are represented by a set of retrieved documents. However, as we would like to evaluate a search engine's performance based on users' information needs satisfaction, it can be argued that we would have to permit users' involvement in the IR process, at least in formulating an information needs statement and evaluating retrieved documents. At the same time, the problem is to capture users' information needs

and analyze the level of their satisfaction without sacrificing the main objective of our experiments: to evaluate a search engine and not the whole IR process.

For this purpose we have to consider the users' involvement in the IR process (Fig.2). At the beginning, a user recognizes an information need, tries to formulate an information needs statement, and to translate it into a query. A query is composed based on the user's knowledge of the constraints of the search engine (for example, indexing algorithm), as well as on the flexibility of the interface between a user and the search engine. When IR results are presented through an interface (or in printed form), a user evaluates a document's relevance to the query based on her knowledge of a subject and an understanding of the information need.

Users can become involved in the evaluation process in the following ways (Fig.2):

1. If we permit users to participate in all stages of the process outlined above, we can gather useful information about user participation in the IR process; for example: query formulation, interaction with a search engine, and evaluation of IR results. However, all these make a user an active participant in the IR process. Therefore, any such evaluation would be an evaluation of the IR process, rather than the search engine, which is our primary goal.

2. As an alternative, we can restrict a user to the evaluation of IR results. One of the scenarios of this experiment would be the following. All components of the IR process are kept "constant," except the settings of a search engine. A query is presented to a user together with IR results (a set of retrieved documents) in a printed form or on the screen. The user's task is to make a decision about the relevance of retrieved documents to the query. The process of decision-making is profiled using methods of cognitive psychology: "think aloud" protocol analysis (Ericsson and Simon 1993; Kushniruk, Kaufman et al. 1996; Kushniruk and Patel 1998) and analysis of user-system interaction using, for example, online monitoring methods (Borgman, Hirsh et al. 1996).

On the surface, the last outlined method can be used for the evaluation of user satisfaction with IR results obtained by search engines with different settings. However, users' characteristics can vary considerably between experiments. Moreover, users' decisions about relevance are based on too many different factors of an IR process. Therefore, the outlined method requires active participation of a user and evaluates the whole IR process rather than just a search engine. At the same time, this method can be used in an analysis of the dynamics of the decision-making process of different categories of users about retrieved document(s).

I argue that, for the specific evaluation of a search engine's ability to satisfy users' information needs, a user has to be prohibited from participating in the IR process. This means that a user neither knows about a problem, nor formulates INS, composes a query, or evaluates retrieved documents. How, then, is it possible to know users' information needs, if a user is not permitted to participate in the experiment? The key point of my argument is that users are allowed to participate in evaluation experiments, but not in the IR process.

I argue that the crucial point lies in the relationship between a user and a query. Until now, IR evaluation experiments involving a "relevance judgement" were based on a decision made when a user knew a query. The knowledge of the query introduces the binary "relevance judgement": a document is relevant to a query or not. Therefore, I argue that if a user neither creates a query, nor knows about it during an evaluation of retrieved documents, it would help to create methods for an external evaluation of search engines.

One alternative would be to use a randomly selected query composed by an expert. This query is submitted to a search engine and the semantics of retrieved document(s) are compared to a query's semantics. The assumption is that a query is a reliable representation of users' information needs. In this case, a user does not participate in an evaluation. Therefore, this experiment can be used for an evaluation of a search engine's ability to satisfy users' information needs.

This experimental design is used in cases when users' information needs are expressed in the form of a query (methods A, B, and C)

Another approach is needed to evaluate satisfaction of users' information needs when they are in the form of INS and ASK (methods D, E, F, G, H, and I). I propose that it can be done using the following experimental design: a user who does not know about a query is presented with retrieval results and asked to decide which information needs statements could be satisfied by such results. In other words, we ask a user the following question: "If these retrieved document(s) are answers, what could be the question(s)?" I call this type of experiment a "jeopardy game," based on the famous TV show.

There can be different designs of these experiments. A user can be asked either to produce a possible main INS (method D) or different possible INSs (methods E and F). Another possibility is to ask a user to "think aloud" during this experiment to gather information about possible INS and ASK for methods D, E, F, G, H, and I. An analysis of a "think aloud" protocol would facilitate an understanding of the many ways users work with the cognitive structure constituting their ASK. Researchers could also capture users' information needs stated in a natural language. Based on this information, we can formally represent potential users' information needs.

One such experiment could be:

1. Participants are independently presented with IR results. The task is to name possible information needs statement(s) that a particular document may be relevant to. (A decision has to be made by the experimenter about the possible number of such statements; for example: would it be limited or not, would a time limit be imposed on a user with each document and with the whole retrieved set.)

2. All produced statements are analyzed with regard to each document, and the whole set. This could be done using methods of "conceptual" overlap. Some potential inferences might be: the scope and distribution of information needs that a specific document can satisfy, and the scope and distribution of information needs that the whole retrieved set can satisfy. By distribution, we understand the relative "weight" of each information need, which is assigned based on frequency analysis of all information needs statements specified by the participants.

Comparison of search engines

Methods C, F, and I address the comparison of search engines based on users' information needs satisfaction. Method C compares retrieved sets of documents when users' information needs are represented by a query. Methods F and I compare the performance of search engines when users' information needs are represented by an information need statement and an ASK correspondingly.

In an attempt to answer question C, one possibility is to compare the documents from different retrieved sets and determine their overlap. However, a simple imaginary experiment shows that this approach is faulty. Let's assume that from a pool of 50 documents, which can be considered relevant to a query, each one of two search engines retrieves different sets of 20 documents. These sets can be compared based on the number of documents common in both sets. I call this type of overlap "physical." In this imaginary experiment, the "physical" overlap is zero. However, both sets of documents come from the same relevance base. Therefore, all of these documents share similar "semantic space(s)" and their "conceptual" overlap can be as high as 100%.

An alternative is to measure the semantic content of every document in some fashion and to use this as the basis for comparison. The idea of "conceptual" overlap is based on the fact that if documents can be judged as relevant to a query, they have to have similar "semantic spaces" and share some concepts and

relationships. In order to capture this "conceptual" overlap, it is possible to resort to a formal representation and compare query and document information. (Some approaches to the practical realization of this idea will be presented later.) However, a method of comparison between formal representations of the semantics of two (or more) sets of retrieved documents has to be further investigated.

Possible approaches to the capturing and comparison of semantics

In this thesis I will only briefly outline how "conceptual" overlap could be evaluated. There are many problems that have to be solved before such an evaluation can be widely used. To be able to show "conceptual" overlap we have first to use a method that permits us to capture and represent formally the semantics of text. Ideally, the chosen method has to be able to capture the semantics of text from different processes: documents in a document set, queries submitted to a search engine, and transcripts of "think aloud" protocol capturing user interaction with retrieved documents (for example, evaluation of results, clarification of an information request, change of a query, and others).

The main purpose of capturing semantics is to achieve a formalized representation of a meaning. This is very similar to a process of concept representation. There are different approaches to capturing semantics. Some of these methods are used in artificial intelligence, and others are from cognitive

psychology. The most often used methods in artificial intelligence are frames (Minsky 1981), semantic networks (Quillian 1968; Collins and Quillian 1969; Brachman 1979; Brachman and Levesque 1985; Sowa 1991), and conceptual graphs (Sowa 1984). In cognitive psychology, researchers use methods of propositional analysis (Kintsch 1985; Kintsch 1989; Kintsch 1998), prose analysis (Meyer 1985), Frederiksen's grammar (Frederiksen 1975; Frederiksen 1985). We compared methods of medical concepts representation used: natural language processing, controlled terminologies, knowledge-based approaches, and statistical representations (Kagolovsky, Miller et al. 1997). Kintsch (Kintsch 1989) has offered a detailed comparison of different methods of semantics capturing (features systems, associative networks, semantic networks, schemes, frames, scripts, and propositional analysis). This comparison shows that propositional analysis is a robust, flexible method that is best suited for analysis of cognitive processes. Based on this comparison, and on the fact that in IR we deal with analysis of cognitive processes either directly (a user) or indirectly (text), I have chosen Kintsch's propositional analysis as a method for a future research.

Capturing the semantics of a query and INS is much easier than capturing the semantics of ASK, since queries and INS are expressed as a text, and ASK are cognitive structures. One possibility could be to present users with a problem and to ask them to "think aloud" about it. Analysis of the protocol can reveal users' knowledge "gaps" related to the problem. These "gaps" represent users'

ASK. They could be represented formally and compared with the meaning of retrieved document(s).

Semantics comparison is a more complicated issue that has to be researched further. However, as results of propositional analysis can be represented as a network of propositions, it is possible to explore graph-theoretic methods (Sowa 1984), overlay analysis (Carr and Goldstein 1977), and other strategies. Kintsch (Kintsch 1989; Kintsch 1998) also differentiates between the absolute meaning of a text - e.g., as intended by the author - and the meaning that a user extracts from the text under certain circumstances. I propose using this idea to differentiate between information need and query by capturing both with different experimental approaches: the absolute semantics through propositional analysis, or related methods, and the information needs through analysis of the cognitive approach of users to understanding text. Once this is achieved, a major stumbling block to understanding the static (and eventually dynamic) relations of the semantics of document contents, queries and information needs, and therefore to their comparison, is removed. This understanding is a prerequisite for improving information retrieval by creating better search engines capable of achieving better satisfaction of user information needs.

Difference between the proposed and currently used evaluation methods

The proposed methods are different from the widely-used measures of recall and precision because they:

1. Evaluate search engines specifically.

2. Permit a more versatile and specific use of the concept of "relevance" and are grounded in a modern understanding of the cognitive processes involved in relevance judgement.

3. Provide a link between the functionality of search engines and users' information needs.

The method of "think-aloud" protocol analysis is used in cognitive psychology to evaluate users' interaction with information systems. Users' interaction with a system is recorded (sound and video) and analyzed. Although this method provides useful information for the improvement of information systems functionality and users' satisfaction, it cannot solve the main problems of relevance-based evaluation. I intend to prove this in the future by comparing the results received using new proposed methods with results of "think-aloud" protocol analysis.

These are some differences between these approaches:

1. "Think aloud" protocol analysis focuses on users' cognitive processes, for example, during a user-system interaction. The proposed methods provide direct evaluation of a search engine's functionality.

2. "Think aloud" protocol analysis evaluates the interaction between users and a system. If our methods allow us to capture the absolute meaning of documents, we will be able to assess systems even when there is no a user-system interaction.

3. "Think aloud" protocol analysis evaluates the process of relevance judgement. The proposed methods permit relevance judgement and the functionality of a search engine to be linked.

CHAPTER 8: CONCLUSIONS

Summary of the thesis

My thesis has demonstrated that the systems analytic approach can help in improving the evaluation of information retrieval. Based on the literature review of information retrieval in general, and its evaluation in particular, problems were identified and solutions to improving IR evaluation were proposed. New methodologies for a comprehensive evaluation of search engines and "relevance" relationships have been proposed, and their comparison with the existing evaluation approaches has been presented.

Contributions

The following contributions to IR research can be identified as the results of this thesis:

1. A new comprehensive definition of information retrieval has been proposed.
2. Two new models of IR components and the IR process have been created.
3. A theoretical framework for improving analysis of the IR process has been presented.
4. Based on this framework, the boundaries of IR components have been identified, and a basis for an improved terminology of IR has been introduced.

5. This is the first time that it has been demonstrated that relevance based methods of evaluation of recall and precision do not evaluate search engines, but rather the whole IR process.
6. New approaches to a comprehensive evaluation of search engines has been proposed.
7. A more comprehensive and precise approach to identifying different kinds of "relevance" relationships in the IR process has been proposed.
8. A theoretical basis for improving evaluation of "relevance" relationships has been identified.

Future research

There is still a lot of work that needs to be done in improving evaluation of information retrieval. I have identified the following directions of future research:

1. Create a method of evaluating semantic overlap between texts.

As was shown in this thesis, we need to use methods of cognitive psychology toward a better evaluation of IR. In particular, there is a need for a simple, consistent and reliable method of capturing text semantics. Such methods go

beyond decisions that are currently made regarding relevance judgements. Text semantics is a complex structure. Evaluating the overlap between formally represented cognitive structures can be a challenge. Some examples given in this thesis demonstrate the complexity of this problem.

2. Classification of evaluation methods used in IR.

There are currently a large number of different approaches to the evaluation of information retrieval. Authors often use evaluation methods based on their preferences and research backgrounds. A classification of currently available evaluation methods can be useful for researchers. A systems analytic approach and, especially, proposed models of IR, provide a firm basis for planning evaluation, choosing methods, and comparing results. This classification would be based on a comprehensive literature review and analysis.

3. Creation of a comprehensive evaluation framework to guide selection, research and design of information retrieval systems.

This framework would evaluate the entire IR process, as well as its individual components, for example, search engine, computer-user interface. This evaluative framework would be applied to the comparison of existing search engines, and more precisely to evaluating differences in IR performance, to

achieve better understanding of the relationship between search engine functionality and the satisfaction of users' information needs.

REFERENCES

Allen, B. (1991). Cognitive Research in Information Science: Implications for Design. Annual Review of Information Science and Technology. M. E. Williams. Medford, NJ, Learned Information Inc. for the American Society for Information Science (ASIS). **26:** 3-37.

Anderson, J. G., C. E. Aydin, et al., Eds. (1994). Evaluating Health Care Information Systems: Approaches and Applications. Thousand Oaks, CA, Sage.

Barry, C. L. (1994). "User-Defined Relevance Criteria: An Exploratory Study." Journal of the American Society for Information Science **45**(3): 149-159.

Beaulieu, M., S. Robertson, et al. (1996). "Evaluating Interactive Systems in TREC." Journal of the American Society for Information Science **47**(1): 85-94.

Belkin, N. J. (1984). "Cognitive models and information retrieval." Social Science Information Studies **4**: 111-129.

Belkin, N. J., R. N. Oddy, et al. (1982). "ASK for information retrieval: Part I. Background and theory. Part II. Results of design study." Journal of Documentation **38**: 61-71, 145-164.

Blair, D. C. (1996). "STAIRS Redux: Thoughts on the STAIRS Evaluation, Ten Years after." Journal of the American Society for Information Science **47**(1): 4-22.

Blair, D. C. and M. E. Maron (1985). "An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System." Communications of the ACM **28**: 289-299.

Blair, D. C. and M. E. Maron (1990). "Full-text Information Retrieval: Further Analysis and Clarification." Information Processing & Management **26**(2): 437-447.

Blauberg, I. V. and V. N. Sadovsky (1977). Systems Theory: Philosophical and Methodological Problems. Moscow, Progress Publishers.

Borgman, C. L., S. G. Hirsh, et al. (1996). "Rethinking Online Monitoring Methods for Information Retrieval Systems: From Search Product to Search Process." Journal of the American Society for Information Science **47**(7): 568-583.

Bowler, D. T. (1981). General Systems Thinking: Its Scope and Applicability. New York, NY, Elsevier North Holland, Inc.

Brachman, R. J. (1979). On the epistemological status of semantic networks. Associative Networks: Representation and Use of Knowledge by Computers. N. V. Findler. New York, NY, Academic Press**:** 3-50.

Brachman, R. J. and H. J. Levesque, Eds. (1985). Readings in knowledge representation. Los Altos, CA, Morgan Kaufmann Publishers, Inc.

Bruce, H. W. (1994). "A Cognitive View of the Situational Dynamism of User-Centered Relevance Estimation." Journal of the American Society for Information Science **45**(3): 142-148.

Buckland, M. and F. Gey (1994). "The Relationship between Recall and Precision." Journal of the American Society for Information Science **45**(1): 12-19.

Buckland, M. and C. Plaunt (1994). "On the Construction of Selection Systems." Library Hi Tech **4**(12): 15-28.

Carr, B. and I. P. Goldstein (1977). Overlays: a theory of modeling for computer-assisted instruction. Cambridge, MA, Massachusetts Institute of Technology.

Checkland, P. (1981). Systems thinking, systems practice. New York, J. Wiley.

Cleverdon, C. (1967). The Cranfield tests on index language devices. Aslib Proceedings.

Cleverdon, C. W. (1970). "Evaluation Tests of Information Retrieval Systems." Journal of Documentation **26**(1): 55-67.

Cleverdon, C. W. (1991). The significance of the Cranfield tests on index languages. ACM SIGIR'91.

Cleverdon, C. W. and E. M. Keen (1966). Factors Determining the Performance of Indexing Systems. Cranfield, UK, Aslib Cranfiled Research Project.

Cleverdon, C. W. and R. G. Thorne (1954). "An Experiment with the Uniterm System." RAE Library Memo(7).

Collins, A. M. and M. R. Quillian (1969). "Retrieval from semantic memory." Journal of Verbal Learning and Verbal Behavior **8**: 240-247.

Cooper, W. S. (1968). "Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems." American Documentation **19**: 30-41.

Cooper, W. S. (1973). "On Selecting a Measure of Retrieval Effectiveness." Journal of the American Society for Information Science **24**: 87-100.

Cooper, W. S. (1976). "The Paradoxical Role of Unexamined Documents in the Evaluation of Retrieval Effectiveness." Information Processing & Management **12**: 367-375.

Croft, W. B. (1993). "Knowledge-based and statistical approaches to text retrieval." IEEE Expert **8**: 8-12.

Cuadra, C. A. and R. V. Katter (1967). Experimental Studies of Relevance Judgments. Volume I: Project Summary. Santa Monica, CA, System Development Corp.**:** 129 p.

Cuadra, C. A. and R. V. Katter (1967). "Opening the Black Box of "Relevance"." Journal of Documentation **23**(4): 291-303.

Czaja, S. J. (1997). Systems Design and Evaluation. Handbook of human factors and ergonomics. G. Salvendy. New York, NY, John Wiley & Sons, Inc.**:** 17-40.

Daniels, P. J. (1986). "Cognitive Models in Information Retrieval: An Evaluative Review." Journal of Documentation **42**: 272-304.

Dervin, B. and M. S. Nilan (1986). Information Needs and Uses. Annual Review of Information Science and Technology. M. E. Williams. White Plains, NY,

Knowledge Industry Publications, Inc. for the American Society for Information Science. **21:** 3-33.

Egan, D. E., J. R. Remde, et al. (1989). "Formative Design-Evaluation of SuperBook." ACM Transactions on Information Systems **7**(1): 30-57.

Eisenberg, M. B. and L. Schamber (1988). Relevance: The Search for a Definition. ASIS Proceedings. Atlanta, GA**:** 164-168.

Ellis, D. (1996). "The Dilemma of Measurement in Information Retrieval Research." Journal of the American Society for Information Science **47**(1): 23-36.

Ericsson, A. and H. A. Simon (1993). Protocol analysis: Verbal reports as data. Cambridge, MA, MIT Press.

Fairthorne, R. A. (1964). Basic Parameters of Retrieval Tests. Proceedings of the 1964 Annual Meeting of the American Documentation Institute, Washington, DC, Spartan Books.

Fidel, R. (1993). "Qualitative methods in information retrieval research." LISR **15**: 219-247.

Fidel, R. and M. Crandall (1997). "Users' Perception of the Performance of a Filtering System." SIGIR: 198-205.

Fidel, R. and D. Soergel (1983). "Factors Affecting Online Bibliographic Retrieval: A Conceptual Framework for Research." Journal of the American Society for Information Science **34**(3): 163-180.

Foskett, D. J. (1972). "A Note on the Concept of "Relevance"." Information Storage and Retrieval **8**(2): 77-78.

Frederiksen, C. H. (1975). "Representing logical and semantic structure on knowledge acquired from discourse." Cognitive psychology **7**: 371-458.

Frederiksen, C. H. (1985). Cognitive models and discourse analysis. Written communication annual: an international survey of research and theory (studying writing: linguistics approaches). C. R. Cooper and S. Greenbaum. Beverly Hills, CA, Sage. **1**.

Friedman, C. P. and J. C. Wyatt (1997). Evaluation Methods in Medical Informatics. New York, NY, Springer-Verlag Inc.

Froehlich, T. J. (1994). "Relevance Reconsidered - Towards an Agenda for the 21st Century: Introduction to Special Topic Issue on Relevance Research." Journal of the American Society for Information Science **45**(3): 124-134.

Greenberg, I. and L. Garber (1999). "Searching for New Search Technologies." Computer(August): 4, 6-7, 11.

Group, O. S., Ed. (1981). Systems Behavior. London, UK, Paul Chapman Publishing, Ltd.

Guba, E. G. and Y. S. Lincoln (1981). Effective Evaluation. San Francisco, CA, Jossey-Bass.

Harman, D. (1995). "Overview of the Second Text REtrieval Conference (TREC-2)." Information Processing & Management **31**(3): 271-289.

Harter, S. P. (1971). "The Cranfield II Relevance Assessments: A Critical Evaluation." Library Quarterly **41**: 229-243.

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

Harter, S. P. (1996). "Variations in Relevance Asessments and the Measurement of Retrieval Effectiveness." Journal of the American Society for Information Science **47**(1): 37-49.

Harter, S. P. and C. A. Hert (1997). Evaluation of Information Retrieval Systems: Approaches, Issues, and Methods. Annual Review of Information Science and Technology (ARIST). M. E. Williams. Medford, NJ, Information Today, Inc. **32:** 3-94.

Hernon, P. and C. R. McClure (1990). Evaluation and Library Decision Making. Norwood, NJ, Ablex Publishing Co.

Hersh, W. (1994). "Relevance and Retrieval Evaluation: Perspectives from Medicine." Journal of the American Society for Information Science **45**(3): 201-206.

Hersh, W., J. Pentecost, et al. (1996). "A Task-Oriented Approach to Information Retrieval Evaluation." Journal of the American Society for Information Science **47**(1): 50-56.

Hersh, W. R. (1996). Information Retrieval: A Health Care Perspective. New York, N.Y., Springer-Verlag New York, Inc.

Hersh, W. R., D. L. Elliot, et al. (1994). Towards New Measures of Information Retrieval Evaluation. Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, Philadelphia, Hanley & Belfus Inc.

Hert, C. A. (1997). Understanding Information Retrieval Interactions. Greenwich, CT, Ablex Publishing Corp.

House, E. R. (1980). Evaluating with Validity. Beverly Hills, CA, Sage.

Howard, D. L. (1994). "Pertinence as Reflected in Personal Constructs." Journal of the American Society for Information Science **45**(3): 172-185.

Hull, D. A. (1996). "Stemming Algorithms: A Case Study for Detailed Evaluation." Journal of the American Society for Information Science **47**(1): 70-84.

Kagolovsky, Y., M. Miller, et al. (1997). Statistical concept representation for indexing of clinical narratives. COACH Conference 22, Vancouver, BC, Canada, HC&CC (Healthcare Computing & communications Canada, Inc.), Edmonton, Alberta, Canada.

Kantor, P. B. (1994). Information Retrieval Techniques. Annual Review of Information Science and Technology. M. E. Williams. Medford, N.J., American Society for Information Science (ASIS). **29:** 53-90.

Kemp, D. A. (1974). "Relevance, Pertinence and Information System Development." Information Storage and Retrieval **10**(2): 37-47.

Kent, A., M. M. Berry, et al. (1955). "Machine literature searching. VIII: Operational criteria for designing information retrieval systems." American Documentation **6**: 93-101.

Kintsch, W. (1985). Text processing: A psychological model. Handbook of Discourse Analysis: Dimensions of Discourse. London, Academic Press. **2:** 231-243.

Kintsch, W. (1989). The representation of knowledge and the use of knowledge in discourse comprehension. Language Processing in Social Context. R. Dietrich and C. F. Graumann, Elsevier Science Publishers B.V. (North-Holland)**:** 185-209.

Kintsch, W. (1998). Comprehension: A paradigm for cognition. New York, Cambridge University Press.

Kushniruk, A. W., D. R. Kaufman, et al. (1996). "Assessment of a computerized patient record system: A cognitive approach to evaluating medical technology." M.D. Computing **13**(5): 406-415.

Kushniruk, A. W. and V. L. Patel (1998). "Cognitive evaluation of decision making processes and assessment of information technology in medicine." Int. J. Med. Inform. **51**: 83-90.

Lancaster, F. W. and A. J. Warner (1993). Information Retrieval Today. Arlington, VA, Information Resources Press.

Lantz, B. E. (1981). "The Relationship between Documents Read and Relevant References Retrieved as Effectiveness Measures for Information Retrieval Systems." Journal of Documentation **37**: 134-145.

Lesk, M. E. and G. Salton (1969). "Relevance Assessments and Retrieval System Evaluation." Information Storage and Retrieval **4**: 343-359.

Losee, R. M. (1996). "Evaluating Retrieval Performance Given Database and Query Characteristics: Analytic Determination of Performance Surfaces." Journal of the American Society for Information Science **47**(1): 95-105.

Losee, R. M. (1997). "Comparing Boolean and Probabilistic Information Retrieval Systems across Queries and Disciplines." Journal of the American Society for Information Science **48**(2): 143-156.

Mannoni, B. (1996). "Bringing museums online." Communications of the ACM **39**(6): 100-105.

Marchionini, G. (1992). "Interfaces for End-User Information Seeking." Journal of the American Society for Information Science **43**(2): 156-163.

Meadow, C. T. (1985). "'Relevance?'." Journal of the American Society for Information Science **36**: 354 -355.

Meadow, C. T. (1986). "Problems of information science research - an opinion paper." Canadian Journal of Information Science **11**: 18-23.

Meadow, C. T. (1992). Text Information Retrieval Systems. San Diego, Academic Press, Inc.

Mesarovic, M. D. and Y. Takahara (1989). Abstract Systems Theory. Berlin, Heidelberg, Springer-Verlag.

Meyer, B. J. F. (1985). Prose Analysis: Purposes, Procedures, and Problems. Understanding expository text. B. K. Britton and J. B. Black. Hillsdale, NJ, Erlbaum: 11-64.

Minker, J. (1977). "Information Storage and Retrieval - A Survey and Functional Description." SIGIR Forum, Association for Computing Machinery **12**(2): 1-108.

Minsky, M. (1981). A framework for represening knowledge. Mind Design. J. Haugeland. Cambridge, MA, MIT Press: 95-128.

Mizzaro, S. (1997). "Relevance: The Whole History." Journal of the American Society for Information Science **48**(9): 810-832.

Nilan, M. S., R. P. Peek, et al. (1988). A Methodology for Tapping User Evaluation Behaviors: An Exploration of Users' Strategy, Source and Information Evaluating. ASIS '88: Proceedings of the American Society for Information Science (ASIS) 51st Annual Meeting, Atlanta, GA, Medford, NJ: Learned Information, Inc. for the American Society for Information Science.

Ogden, C. K. and I. A. Richards (1946). The Meaning of Meaning. New York, Harcourt, Brace and World.

Park, T. K. (1993). "The nature of relevance in information retrieval: An empirical study." Library Quarterly **63**: 318-351.

Park, T. K. (1994). "Toward a Theory of User-Based Relevance: A Call for a New Paradigm of Inquiry." Journal of the American Society for Information Science **45**(3): 135-141.

Quillian, M. R. (1968). Semantic Memory. Semantic Information Processing. M. Minsky. Cambridge, MA, MIT Press.

Raghavan, V. V., G. S. Jung, et al. (1989). "A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance." ACM Transactions on Information Systems **7**(3): 205-229.

Rapoport, A. (1986). General System Theory: Essential Conepts and Applications. Cambridge, Mass, Abascus Press.

Rees, A. M. and D. G. Schultz (1967). A Field Experimental Approach to the Study of Relevvance Assessments in Relation to Document Searching. Volume I. Cleveland, OH, Case Western Reserve University, School of Library Science, Center for Documentation and Communication Research: 305 p.

Robertson, S. E. and M. M. Hancock-Beaulieu (1992). "On the Evaluation of IR Systems." Information Processing & Management **28**(4): 457-466.

Rossi, P. H. and H. E. Freeman (1989). Evaluation: A Systematic Approach. Newbury Park, CA, Sage.

Salton, G. (1986). "Another look at automatic test-retrieval systems." Communications of the ACM **20**: 648-656.

Salton, G. (1992). "The State of Retrieval System Evaluation." Information Processing & Management **28**(4): 441-449.

Salton, G. and M. E. Lesk (1968). "Computer Evaluation of Indexing and Text Processing." Journal of the ACM **15**(1): 8-36.

Salton, G. and M. J. McGill (1983). Introduction to Modern Information Retrieval. New York, McGraw-Hill.

Salvendy, G., Ed. (1997). Handbook of human factors and ergonomics. New York, NY, John Wiley & Sons, Inc.

Saracevic, T. (1970). The Concept of "Relevance" in Information Science: A Historical Review. Introduction to Information Science. T. Saracevic. New York, NY, R.R. Bowker**:** 111-151.

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

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

Schamber, L., M. B. Eisenberg, et al. (1990). "A re-examination of relevance: Toward a dynamic, situational definition." Information Processing & Management **26**: 755-776.

Schatz, B. and H. Chen (1999). "Digital Libraries: Technological Advances and Social Impacts." Computer **32**(2): 45-50.

Scherrer, J.-R. (1998). "Concepts, knowledge and language in healthcare information systems: follow-up 30 months later." Methods Inf Med 1998 Nov;37(4-5):312-4 **37**(4-5): 312-314.

Shaw, D. (1991). The Human-Computer Interface for Information Retrieval. Annual Review of Information Science and Technology. M. E. Williams. Medford, NJ, Learned Information Inc. for the American Society for Information Science (ASIS). **26:** 155-195.

Shaw, W. M., Jr., R. Burgin, et al. (1997). "Performance Standards and Evaluations in IR Test Collections: Vector-Space and Other Retrieval Models." Information Processing & Management **33**(1): 15-36.

Shneiderman, B. (1992). Designing the user interface: strategies for effective human-computer interaction. Reading, MA, Addison-Wesley.

Smeaton, A. (1993). Report on TREC-2 Conference. [online]. IR Digest. **10**.

Sowa, J. F. (1984). Conceptual Structures: Information Processing in Mind and Machines. Reading, MA, Addison-Wesley.

Sowa, J. F., Ed. (1991). Principles of Semantic Networks: Explorations in the Representation of Knowledge. The Morgan Kaufmann Series in Representation and Reasoning. San Mateo, CA, Morgan Kaufmann Publishers, Inc.

Sparck Jones, K. (1995). "Reflections on TREC." Information Processing & Management **31**(3): 291-314.

Su, L. T. (1992). "Evaluation Measures for Interactive Information Retrieval." Information Processing & Management **28**(4): 503-516.

Sugar, W. (1995). User-Centered Perspective of Information Retrieval Research and Analysis Methods. Annual Review of Information Science and Technology. M. E. Williams. Medford, NJ, Information Today Inc. for the American Society for Information Science (ASIS). **30:** 77-109.

Sutton, S. A. (1994). "The Role of Attorney Mental Models of Law in Case Relevance Determinations: An Exploratory Analysis." Journal of the American Society for Information Science **45**(3): 186-200.

Swanson, D. R. (1965). "Evidence Underlying the Cranfield Results." Library Quarterly **35**: 1-20.

Swanson, D. R. (1971). "Some unexplained aspects of the Cranfield tests of indexing performance factors." Library Quarterly **41**: 223-228.

Swanson, D. R. (1977). "Information retrieval as a trial-and-error process." Library Quarterly **47**: 128-148.

Swanson, D. R. (1986). "Subjective versus objective relevance in bibliographic retrieval systems." Library Quarterly **56**: 389-398.

Swanson, D. R. (1988). "Historical note: Information retrieval and the future of an illusion." Journal of the American Society for Information Science **39**: 92-98.

Swets, J. A. (1963). "Information Retrieval Systems." Science **141**: 245-250.

Swets, J. A. (1969). "Effectiveness of Information Retrieval Systems." American Documentation **20**: 72-89.

Tague, J., A. Salminen, et al. (1991). Complete Formal Model for Information Retrieval Systems. ACM SIGIR International Conference on Research and Development in Information Retrieval, Chicago, IL, New York: ACM Press.

Tague, J. and R. Schultz (1989). "Evaluation of the User Interface In An Information Retrieval System: A Model." Information Processing & Management **25**(4): 377-389.

Tague-Sutcliffe, J. M. (1992). "The Pragmatics of Information Retrieval Experimentation, Revisited." Information Processing & Management **28**(4): 467-490.

Tague-Sutcliffe, J. M. (1996). "Some Perspectives on the Evaluation of Information Retrieval Systems." Journal of the American Society for Information Science **47**(1): 1-3.

Vickery, B. and A. Vickery (1993). "Online Search Interface Design." Journal of Documentation **49**: 103-187.

Voiskunskii, V. G. (1997). "Evaluation of Search Results: A New Approach." Journal of the American Society for Information Science **48**(2): 133-142.

von Bertalanffy, L. (1976). General System Theory : Foundations, Development, Applications. New York, NY, George Braziller, Inc.

Voorhees, E. M. (1998). Variations in Relevance Judgments and the Measurement fo Retrieval Effectiveness. ACM SIGIR'98.

Westbrook, L. (1993). "User Needs: A Synthesis and Analysis of Current Theories for the Practitioner." RQ **32**(4): 541-549.

Wildemuth, B. M., R. de Bliek, et al. (1995). "Medical Students' Personal Knowledge, Searching Proficiency, and Database Use in Problem Solving." Journal of the American Society for Information Science **46**(9): 590-607.

Wilson, P. (1973). "Situational relevance." Information Storage and Retrieval **9**: 457-471.

Wilson, T., Ed. (1994). Information Needs and Uses: Fifty Years of Progress?

Wilson, T. D. (1981). "On user studies and information needs." Journal of Documentation **37**: 3-15.

Zobel, J. (1998). How Reliable are the Results of Large-Scale Information Retrieval Experiments? ACM SIGIR'98.

APPENDIX A: GLOSSARY OF SYSTEMS CONCEPTS

Adapted from (Group 1981)

Black box: A (component of a) system that is only considered in terms of its inputs and outputs. Its internal mechanisms are unknown or ignored.

Boundary: The conceptual division between a system and its environment; it may or may not correspond to recognized geographical, physical, legal or cultural divisions and will be drawn according to the observer's purpose.

Environment (of a system): The totality of external conditions and concrete or abstract items which affect the behaviour of a system.

Flow-block diagram: One which displays the subsystems as blocks and the flows between them as arrows. Flows may be of money, material, energy, information or decisions.

Interface: The area between interacting systems or subsystems and the components which overlap.

Process(es): The activities which are undertaken by a system or subsystem.

Structure: The components of a system or subsystem and the relationship between them.

System: A system is an assembly of components, connected together in an organized way. The components are affected by being in the system and the behaviour of the system is changed if they leave it. This organized assembly does something and has been identified as a particular interest.

An assembly of parts connected in an organized way that has been identified by someone as a special interest and that behaves in some way (i.e., does more than just exist).

A structured set of objects and/or attributes, together with the relationship between them.

Systematic: Using a method, or following a plan or an explicit and rational procedure.

Systemic: Using systems ideas; treating things as systems or from a systems viewpoint; pertaining to a system or systems.

VITA

Yuri Kagolovsky, MD

Place of Birth: Dnepropetrovsk, Ukraine

Education

- 1983 Doctor of Medicine (with Distinction) (Dnepropetrovsk, Ukraine)
- 1983-1985 Post-graduate training, Academy of Medical Sciences (Russia)
- 1996-present Graduate student, School of Health Information Science, UVic

Grants/Awards Received

- Best graduate of 1983 (Dnepropetrovsk, Ukraine)
- HEALNet support as Master student (1996-present)
- Prize for student scientific presentation at the HEALNet AGM (Hamilton, 1996)
- Traveling grants to conferences: COACH 1996 (Toronto), AMIA 1996 (Washington, DC), HEALNet AGM (Hamilton, 1996; Toronto, 1997, Calgary, 1999), IMIA WG6 conference 1997 (Jacksonville, FL), COACH 1997 (Vancouver), MedInfo '98 (Seoul, Korea), IMIA WG6 conference 2000 (Phoenix, AZ) (Sources: HEALNet; School of Health Information Science and Faculty of Graduate Studies, University of Victoria)
- President's Research Scholarship (1998/99, 1999/2000, University of Victoria)

Teaching Experience

- Developed and taught tutorials at MedInfo '98 (Seoul, Korea), HEALNet AGM '99 (Calgary, Alberta)
- Developed and taught Health Information Science 270 lab in 1996/97, 1997/98, 1998/99, 1999/2000

Presentations

- Faculty of Medical Informatics, University of Heidelberg / Heilbronn (Heilbronn, Germany, 1999)
- HEALNet NCE mid-term review site visit (Toronto, 1998)
- Student Research Paper Competition (HEALNet AGM 1996, 1997, 1999, 2000)

Memberships

- Student member of HEALNet
- Student member of the American Medical Informatics Association (AMIA)

- Student member of the Medical Informatics Section / Medical Library Association

Research Interests

- Evaluation methodologies in health care
- Information retrieval
- Continuous Quality Improvement (CQI)

Publications

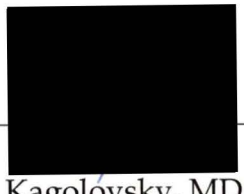
1. Kagolovsky Y and Moehr JR: Evaluation of information retrieval: old problems and new perspectives. In *Proceedings of the 8th International Congress on Medical Librarianship*, London, 2-5 July 2000. Available on the Interenet: <URL: http://www.icml.org/tuesday/ir/kagalovosy.htm>
2. Kagolovsky Y, Freese D, Miller M, Walrod T, Moehr J: Towards improved information retrieval (IR) from medical sources. *International Journal of Medical Informatics*, 51, 1998, pp. 181-195.
3. Kagolovsky Y and Moehr JR: A structured model for evaluation of information retrieval. In *Medinfo '98, Proceedings* (Eds: B. Cesnik et al.), IMIA/IOS Press, Amsterdam, 1998, pp. 171-175.
4. Kagolovsky Y, Miller M and Moehr JR: Statistical concept representation for indexing of clinical narratives. In *COACH Conference 22, Scientific Program Proceedings* (Ed: P. Fisher), Edmonton, 1997, pp. 118-126.
5. Moehr JR, Kagolovsky Y: Integrated health information systems, quality of care and indexing. In *COACH Conference 22, Scientific Program Proceedings* (Ed: P. Fisher), HC&CC, Edmonton, 1997, pp. 87-96.
6. Snisar V, Bujalsky A, Kagolovsky Y et al.: Peculiarities of using a respirator "Mlada" for long-term artificial respiration of newborns. In *5th Conference of Anesthesiologists and ICU Specialists of Ukraine, Proceedings*, Voroshilovgrad, 1988 (publication in Russian)
7. Nepomnyazhikh L, Nepomnyazhikh G, Kagolovsky Y et al.: Morphological study of mucous membranes biopsies in human pathology based on epithelial-connective tissues relationships. In *Bulletin of the Academy of Medical Sciences of the USSR (Siberian Department)*, Novosibirsk, 1984, 4, pp. 57-64. (publication in Russian)
8. Kagolovsky Y, Kamen J and Kamen Y: Mathematical model of a blood sugar regulatory system. In *Collected Work of Young Scientists of the Dnepropetrovsk Medical Institute*, Dnepropetrovsk, 1981. (publication in Russian)

PARTIAL COPYRIGHT LICENSE

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the Library of any other university, or similar institution, on its behalf or for one of its users. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by me or a member of the University designated by me. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Title of Thesis:

A Systems Analytic Approach to the Evaluation of Information Retrieval (IR)

Author: 

Yuri Kagolovsky, MD

September 1st, 2000