

©2009

Ying-Hsang Liu

ALL RIGHTS RESERVED

THE IMPACT OF MESH (MEDICAL SUBJECT HEADINGS) TERMS ON
INFORMATION SEEKING EFFECTIVENESS

by

YING-HSANG LIU

A Dissertation submitted to the
Graduate School-New Brunswick
Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Communication, Information and Library Studies

Written under the direction of

Nina Wacholder

And approved by

New Brunswick, New Jersey

May, 2009

ABSTRACT OF THE DISSERTATION

The Impact of MeSH (Medical Subject Headings) Terms on Information Seeking
Effectiveness

By YING-HSANG LIU

Dissertation Director:

Nina Wacholder

To what extent do MeSH (Medical Subject Headings) terms improve search effectiveness of different kinds of users? We observed four different kinds of searchers using an experimental information retrieval (IR) system: (1) search novices; (2) domain experts; (3) search experts and (4) medical librarians. The information needs were a subset of the relatively difficult topics originally created for the Text REtrieval Conference (TREC). By experimental design, we used 20 search topics in an IR user experiment to alleviate search topic variability. Effectiveness of retrieval was based on the relevance judgments set provided by TREC. Thirty-two participants searched either using a version of the system in which abstracts *and* MeSH terms were displayed or another version in which they had to formulate their own terms based only on the display of abstracts. We found

that MeSH terms were more useful for domain experts than for search experts in terms of the precision measure, even though domain experts did not perceive that MeSH terms were useful. We speculate that because of the technical topics, only the domain experts had the knowledge to understand and therefore make use of the MeSH terms. The primary contributions of this research are: (1) assessment of relative impact of searchers characteristics of domain knowledge and search training on search effectiveness and (2) design and methodology for assessing the usefulness of controlled vocabulary. The effort to create MeSH terms is worthwhile for domain experts' searches on technical topics.

ACKNOWLEDGEMENTS

This study would not have been possible without the advice and support from many individuals. I am deeply grateful to my advisor, Nina Wacholder, for her guidance, encouragement and commitment to her students. I would also like to thank my committee members Michael Lesk and Paul Kantor who provided invaluable feedback to help me develop research ideas and experimental design. I express my gratitude to outside member Chung-chieh Shan who provided feedback on the impact of interface design. I would also like to express thanks to my PhD cohorts, colleagues and faculty in SCILS for a dynamic research environment, especially to Tefko Saracevic, Paul Kantor and Nick Belkin for inspirational seminar courses.

In addition, I would like to acknowledge here those who initiated my interest in research, including Shan-Ju L. Chang and Mei-Mei Wu. It was their teaching and research collaboration that led me to the fields of human information behavior and information retrieval. Their unfailing support has helped me overcome many difficulties.

I would like to express my gratitude to the LIS department for awarding me teaching assistantships and the opportunity for teaching MLIS Organizing Information course. This dissertation research was funded by NSF grant #0414557, PIs. Michael Lesk and Nina Wacholder. I would also like to thank Lu Liu for technical assistance and research participants who shared their expertise.

Finally, I would also like to thank my family who supported me in every way. I am grateful to Yen-Chen Chuang for her companionship and comfort during the writing of this dissertation.

DEDICATION

For My Family

Table of Contents

Abstract	ii
Acknowledgement	iv
Dedication	v
Table of Contents	vi
List of Tables	x
List of Illustrations	xii
CHAPTER 1 INTRODUCTION	1
1.1 PROBLEM STATEMENT	1
1.2 RESEARCH QUESTIONS AND HYPOTHESES	4
1.3 OVERVIEW OF THE STUDY	10
CHAPTER 2 RELATED WORK	13
2.1 ASSESSMENT OF INDEX TERMS	13
2.1.1 <i>Assessment of human-developed index terms</i>	15
2.1.2 <i>Assessment of automatic indexing techniques</i>	18
2.2 QUERY REFORMULATION AND USER PERCEPTION	21
2.2.1 <i>Query reformulation</i>	21
2.2.2 <i>User perceptions of the usefulness of displayed terms</i>	24
2.3 USER CHARACTERISTICS	25
2.3.1 <i>Domain knowledge</i>	26
2.3.2 <i>Search experience</i>	27
CHAPTER 3 RESEARCH METHODOLOGY	30

3.1 OVERVIEW OF METHOD	30
3.1.1 Subjects	31
3.1.2 Experimental design.....	33
3.1.3 Search tasks and incentive system	35
3.1.4 Experimental procedures	36
3.1.5 Experimental system	42
3.1.6 Documents.....	46
3.1.7 Search topics	47
3.1.8 Reliability of relevance judgment sets	48
3.1.9 Limitations of the design	51
3.2 PARTICIPANT CHARACTERISTICS	52
3.2.1 Domain knowledge.....	53
3.2.2 Search training.....	55
3.2.3 Demographic variables.....	57
3.3 DATA ANALYSIS	58
3.3.1 User search performance	58
3.3.2 Query terms characteristics	59
3.3.3 User perceptions and search performance	61
3.4 SUMMARY	62
CHAPTER 4 RESULTS	64
4.1 OVERALL USE OF MESH TERMS	64
4.2 SEARCH EFFICIENCY	65
4.3 SEARCH OUTCOME	68

4.4 COMPARISON OF HUMAN AND COMPUTER PERFORMANCE	73
4.5 QUERY TERMS	77
4.5.1 <i>Number of terms per search session (tokens)</i>	77
4.5.2 <i>Number of unique terms per search session (types)</i>	78
4.5.3 <i>Number of queries per search session</i>	80
4.5.4 <i>Number of terms per query (query length)</i>	81
4.6 QUERY REFORMULATIONS	83
4.7 USER PERCEPTION OF USEFULNESS OF DISPLAYED TERMS AND SEARCH TASK DIFFICULTY	85
4.7.1 <i>Perceived usefulness of displayed terms</i>	85
4.7.2 <i>Perceived search task difficulty</i>	87
4.8 USER PERCEPTION AND OUTCOME MEASURES	89
4.8.1 <i>Usefulness of displayed terms and outcome measures</i>	89
4.8.2 <i>Perceived search task difficulty and outcome measures</i>	92
4.9 USER CHARACTERISTICS	93
4.10 SUMMARY	96
CHAPTER 5 DISCUSSION AND CONCLUSION	98
5.1 PERCEIVED SEARCH TASK DIFFICULTY	98
5.2 PERCEIVED SEARCH TASK DIFFICULTY AND OUTCOME MEASURES	100
5.3 USEFULNESS OF DISPLAYED TERMS AND SEARCH EFFECTIVENESS	101
5.4 DOMAIN KNOWLEDGE	103
5.5 SEARCH TRAINING	105
5.6 SEARCH TOPIC VARIABILITY	106

5.7 EVALUATION OF MESH+ AND MESH– INFORMATION RETRIEVAL SYSTEMS	108
5.8 CONCLUSION	110
5.9 FUTURE RESEARCH	113
Bibliography	115
Appendix A: Consent Form	127
Appendix B: Experimental Guidelines, Sample Document Records and Search Help	129
Appendix C: Pre-Search Searcher Background Questionnaire	139
Appendix D. Post-Search Search Perception and Comprehension Test for Each Search Topic	142
Appendix E. Post-Search Interview Questions	145
Appendix F. 20 Selected Search Topics	146
Curriculum Vita	149

List of Tables

Table 3-1 Distribution of TREC relevance judgments from top 10 documents retrieved by human participants	50
Table 3-2 Correctness of comprehension test by searcher type.....	55
Table 3-3 Amount of MeSH use experience.....	56
Table 3-4 Gender, native language and age by searcher type.....	57
Table 4-1 Use of MeSH terms search field in MeSH+ version	64
Table 4-2 Search effectiveness by searcher types in terms of precision and recall measures.....	68
Table 4-3 Search effectiveness by system version and searcher type in terms of precision and recall	69
Table 4-4 Summary of query terms results.....	83
Table 4-5 Query reformulations by searcher type in terms of the precision measure (two-tailed paired t-test)	84
Table 4-6 Summary of user perception results	89
Table 4-7 Summary of the relation between terms in abstracts usefulness and the outcome measures (N users = 32; N questions = 20; N all searches = 256; statistical significance at 95%).....	90
Table 4-8 Summary of the relation between terms in abstracts usefulness and the precision score by searcher type (N users = 32; N questions = 20; N all searches = 128; statistical significance at 95%)	91

Table 4-9 Summary of the relation between MeSH term usefulness and the outcome measures (N users = 32; N questions = 20; N all searches = 256; statistical significance at 95%).....	91
Table 4-10 Summary of the relation between search task difficulty and the outcome measures (N users = 32; N questions = 20; N all searches = 256; statistical significance at 95%).....	92
Table 4-11 Summary of the relation search task difficulty and the time spent by searcher type (N users = 32; N questions = 20; N all searches = 256; statistical significance at 95%).....	93
Table 4-12 Summary of the relation between user characteristics and the precision score (N users = 32; N questions = 20; N all searches = 256; statistical significance at 95%).....	94
Table 4-13 Summary of the relation between user characteristics and the recall score (N users = 32; N questions = 20; N all searches = 256; statistical significance at 95%).....	95
Table 4-14 Summary of search effectiveness and search efficiency results.....	96

List of Illustrations

Figure 3-1 Four types of searchers categorized by domain knowledge and search training	32
Figure 3-2 4×4 Graeco-Latin square design	34
Figure 3-3 An overview of experimental procedures	36
Figure 3-4 Concept analysis form for training session	38
Figure 3-5 The 2003 MeSH Browser and hierarchical display of the MeSH term hypertension	41
Figure 3-6 Distinction between MeSH+ and MeSH– search interfaces	43
Figure 3-7 MeSH+ version search interface	44
Figure 3-8 MeSH– version search interface	45
Figure 3-9 Sample search topic	47
Figure 3-10 Box plot of undergraduate level biology knowledge by searcher type	53
Figure 3-11 Box plot of graduate level biology knowledge by searcher type	54
Figure 3-12 Box plot of search training by searcher type	56
Figure 3-13 Query log examples	61
Figure 4-1 Histogram of time spent with normal density overlaid by all searches ($N =$ 256)	66
Figure 4-2 Line plot of the mean and standard error of time spent by searcher type and system version	67
Figure 4-3 Line plot of the mean and standard error of square root of precision by searcher type and system version	70
Figure 4-4 Interaction plot between searcher type and system version	72

Figure 4-5 Plot of MAP (mean average precision) by search topic between human and computer searches	74
Figure 4-6 Plot of P10 (precision after 10 documents retrieved) by search topic between human and computer searches	75
Figure 4-7 Plot of P100 (precision after 100 documents retrieved) by search topic between human and computer searches	76
Figure 4-8 Line plot of the mean and standard error of terms per search session by searcher type and system version.....	78
Figure 4-9 Line plot of the mean and standard error of unique terms per search session by searcher type and system version.....	79
Figure 4-10 Line plot of the mean and standard error of queries per search session by searcher type and system version.....	81
Figure 4-11 Line plot of the mean and standard error of terms per query by searcher type and system version	82
Figure 4-12 Line plot of the mean and standard error of perception of displayed terms usefulness by searcher type and system version	86
Figure 4-13 Line plot of the mean and standard error of perceived search task difficulty by searcher type and system version.....	88

CHAPTER 1 INTRODUCTION

Can we yet say that the benefit received is commensurate with the effort of construction? ... Even if controlled retrieval language and thesauri are useful, is their uncontrolled proliferation equally useful? (Vickery, 1970, pp. 136-137)

1.1 Problem Statement

The type and volume of information resources available to users have dramatically increased because of the rapid growth of Internet information services. More importantly, the fact that users are able to access non-library online resources, such as indexed open access journal articles or author's preprint files, has contributed to the decreasing use of online catalogs in recent years. One of the challenges within this changing information landscape is to determine what kinds of information search tools will be most useful for users, and under what conditions these tools can be effective in finding pertinent information.

Given the widespread recognition of the importance of search engines and full text searching, and the time and expense required to create controlled vocabulary, one of the crucial issues is whether human-developed index terms are still needed. Recent reports have suggested reducing or eliminating the use of manual subject indexing because it is not cost effective (Calhoun, 2006) or difficult to understand and use ("Bibliographic Services Task Force," 2005).

Most recently, The Library of Congress Working Group on the Future of Bibliographic Control recommended the re-purposing of LCSH (Library of Congress Subject Headings) and recognition of the potential of computational methods in the

practice of subject analysis, in view of the current economic model for sharing bibliographic records, higher user expectations and rapidly changing information environments (“On the record: Report of the Library of Congress Working Group on the Future of Bibliographic Control,” 2008). The question of the usefulness of human-developed index terms is key to the operation of information retrieval (IR) systems in the networked environment.

Both human-developed index terms and automatic indexing systems are created to assist intended users to resolve their information problems. For ordinary users their information problems are typically represented as short search terms in one single query (see e.g., Spink, Wolfram, Jansen, & Saracevic, 2001) and the vocabulary mismatch between the user queries and the potentially relevant documents has been widely recognized (see e.g., Blair & Maron, 1985; Lancaster, 1969). One possible solution is to identify the user characteristics and their relationship with system features that may contribute to better search performance. However, previous research on the impact of user characteristics in the use of information retrieval systems has suggested that (a) domain knowledge or specific topic knowledge is not correlated with search outcome (Allen, 1991; Pao, Grefsheim, Barclay, Woolliscroft, McQuillan, & Shipman, 1993) and (b) search experience with databases cannot predict search outcome (Fenichel, 1981; Howard, 1982; Sutcliffe, Ennis, & Watkinson, 2000). That raises the question in what specific conditions these system features like human-developed index terms and automatic indexing techniques will be useful for different kinds of users.

From a modern perspective, index terms are a kind of metadata. In recent years many metadata initiatives have been proposed to facilitate information access, retrieval

and discovery due to the availability of digitized resources available on the Web. One of the stated goals of descriptive metadata, in particular, is resource identification and discovery (“Understanding metadata,” 2004). Much effort has been devoted to the indexing and organizing of digitized objects and the development of metadata standards in digital library research (Lesk, 2004). One of the important issues arising from the implementation of metadata schemes in digital library collections is how we can ensure the quality of metadata that is either manually or automatically assigned to the resources. For instance, to guide the development of automatic metadata assignment tools at the INFOMINE project, Paynter (2005) conducted automatic evaluations of a wide range of metadata fields (title, creator, keyphrase, description, LCSH and category) assigned by computer algorithms. This approach implicitly assumes that the assignment of quality metadata will lead to better information access, even though it has not been formally evaluated in an interactive search environment.

From a practical perspective, the research team at the National Science Digital Library (NSDL) found it difficult to develop structured metadata similar to a library union catalog because of the lack of information professionals with adequate subject domain expertise, computer technical expertise and cataloging experiences (Lagoze, Krafft, Cornwell, Dushay, Eckstrom, & Saylor, 2006). It suggests that the implementation of structured metadata in a distributed environment through metadata aggregation requires enormous amount of well-trained personnel support, even though the specification of metadata standards, such as Dublin Core and OAI-PMH (Open Archives Protocol for Metadata Harvesting), is not considered as complex as LCSH.

Both the use of controlled vocabularies, such as LCSH in academic libraries, and the automatic assignment of metadata in digital libraries using information extraction techniques or automatic classification approaches, face the same problem: whether the resource intensive practice of metadata assignment is worthwhile. Further, the usefulness of metadata, either derived directly from the digitized documents or assigned manually or automatically from a list of controlled vocabularies, needs to be evaluated from the perspective of intended users of digital libraries.

1.2 Research Questions and Hypotheses

A review of related work suggests that the conditions under which controlled vocabulary assigned by human indexers will be useful in online database searching are still unclear. More specifically, while these manually created index terms can be very useful for obtaining relevant documents, we have limited knowledge about how displayed index terms can be of most benefit to search results enhancement in an interactive search environment. Several related threads of research have contributed to our understanding of different components of the whole interactive search process, but each focuses on particular interest of concern within the research paradigm.

In information retrieval research, the evaluation experiments have focused on the retrieval effectiveness of different search algorithms rather than on the effect of different controlled indexing languages (Sparck Jones, 2005). While approaches such as interactive query expansion and relevance feedback have been proposed to address the issue of short user queries, these studies have focused on the use of automatic indexing techniques, rather than human-developed index terms (e.g., Belkin et al., 2000; Efthimiadis, 1996; Ruthven, 2003). Studies of end-user searching have suggested that

domain expert search strategies may be helpful for domain novices when they attempt to do subject searches (see e.g., Markey, 2007a, b). One of the outstanding questions is whether searchers with different levels of domain knowledge and search experience are able to recognize useful terms for query reformulations and obtain better search results.

In this study we use MeSH (Medical Subject Headings) terms as a case study to examine the usefulness of human-developed index terms. MeSH terms are widely recognized as one of the most sophisticated and start-of-the-art controlled vocabularies for IR systems (e.g., Aronson, Bodenreider, Chang, Humphrey, Mork, Nelson, et al., 2000; Nelson, Johnston, & Humphreys, 2001; Humphrey, Rogers, Kilicoglu, Demner-Fushman, & Rindflesch, 2006). MeSH terms are maintained by professional indexers in the National Library of Medicine and used to index bibliographic records in MEDLINE database. The large and high-quality manually indexed documents in MEDLINE database have been critical to biomedical research and development.

The usefulness of MeSH in biomedical searching is especially important because of the extreme popularity of the PubMed database, the public accessible version of MEDLINE on the Web. There are approximately 2.4 million PubMed user searches per day in 2006 based on the PubMed statistics by National Center for Biotechnology Information (NCBI) ("PubMed searches," 2007).

The present study is designed to assess the usefulness of MeSH in searching for biomedical information. A review of related research literature has shown that the research questions asked are germane to several areas of inquires that are concerned about different aspects of the question. The research questions are pursued within different branches of the field of library and information science. These include

organization of information, query formulation in information retrieval and online search behavior studies.

For researchers and practitioners working in the area of organization of information, descriptors, subject heading lists and metadata schema have been commonly used to organize information objects for information access, retrieval and discovery (e.g., Anderson & Perez-Carballo, 2005; Taylor & Joudrey, 2008). Many information retrieval researchers believe that the automatic construction of full-document indexes using various retrieval techniques, without massive efforts of manual indexing, is adequate for retrieving relevant information (e.g., Croft, 1989; Salton, 1986). However, these studies have generally been conducted in a laboratory setting. Online interactive search behavior studies have focused on the impact of searchers' characteristics on search outcomes and search processes (e.g., Barry et al., 2005) in order to equip information professionals with better search skills and ultimately better information services for end-users.

This study aims to provide answers to the question: to what extent do MeSH terms improve search effectiveness for different kinds of users? More specifically, our goal is to determine whether human-developed index terms help users by examining the following research questions:

1. Do MeSH terms help users overall?
2. Do MeSH terms help different kinds of users?
3. What is the relationship between the searcher's perceptions of the usefulness of MeSH terms and their search performance?

As an initial step to answering the overarching question of whether human-developed index terms help users, we formulate the following research hypotheses:

H1. Queries searched using MeSH will get better results than queries searched not using MeSH.

We assess quality of terms by measuring quality of search results in terms of search effectiveness. In IR experiments the search effectiveness of different retrieval techniques is achieved by comparing the search performance of queries. IR researchers have widely used the micro-averaging method of performing statistics on the queries in summarizing precision and recall values for comparing the search effectiveness of different retrieval techniques in order to meet the statistical requirements (see e.g., Tague, 1981; Tague-Sutcliffe, 1992; van Rijsbergen, 1976). The method of micro-averaging is intended to obtain reliable results in comparing search performance of different retrieval techniques by giving equal weights to each query. This hypothesis is concerned with experimental considerations, specifically for robust statistical design, rather than whether MeSH terms will be helpful for a random query.

Within an interactive IR experiment environment that involves human searchers, it is difficult to use a large set of search topics. Empirical evidence has demonstrated that the search topic set size of 50 is necessary to determine the relative performance of different retrieval techniques in batch mode evaluations (Buckley & Voorhees, 2005), because the variability of search topics has an overriding effect on search results. This problem has been exacerbated by the fact that we have little understanding about the nature and properties of search topics for evaluation purposes (Robertson, 1981). More specifically, it is still elusive what kinds of search topics can be used to directly control the topic effect for IR evaluation purposes. Nonetheless, this study has explicitly

diminished the overriding topic effect by an experimental design that controls searchers, systems and search topic pairs and uses a relatively large number of search topics.

The second hypothesis is concerned with the usefulness of MeSH terms at the level of searchers:

H2. Searchers using MeSH will get better results than searchers who do not use MeSH.

In addition to the experimental considerations, we are concerned about the impact of searcher characteristics on search effectiveness for the use of MeSH terms. Previous research on impact of searcher characteristics, particularly domain knowledge and search experience, on search effectiveness has been inconclusive (Fenichel, 1981; Howard, 1982; Hsieh-Yee, 1993; McKibbin, Haynes, Walker Dilks, Ramsden, Ryan, Baker, L., et al., 1990; Pao et al., 1993). It is still unclear for which kinds of searchers the MeSH terms will be most useful in the context of searching complex biomedical topics.

The third set of hypotheses proposes that the answers to question H2 depends on the characteristics of the searcher:

H3. Quality of search results using MeSH will vary by user type.

H3a. Domain experts using MeSH will get better results than domain novices using MeSH.

H3b. Search experts using MeSH will get better results than untrained searchers using MeSH.

Since domain experts can better understand relatively technical biomedical topics and MeSH terms, they are expected to obtain better search results. Search experts are expected to do better than untrained searchers because they are familiar with controlled

vocabularies like MeSH terms and system features in their interactions with an informational retrieval system.

This set of hypotheses should be considered exploratory rather than seeking significant differences, because previous investigations have suggested large individual differences (e.g., Bellardo, 1985; Fenichel, 1981; Saracevic, 1991; Saracevic & Kantor, 1988) and relatively small differences in system performance (see e.g., Sparck Jones, 1974, 1981). Within these constraints, it is potentially difficult to identify statistically significant differences in search results. Nonetheless, the exploratory investigation will contribute to our understanding of the impact of user characteristics and search assistance tools on information seeking effectiveness.

Searcher's perception of the search process is another important aspect of query reformulation tasks in an interactive search environment. To make search tools useful for complex query reformulation tasks, one of the fundamental questions is whether users perceive the search terms suggested by systems useful for their current tasks, and whether users are able to recognize potentially useful terms and ultimately obtain better search performance.

Prior research has suggested that domain expert's selection of expanded terms from a thesaurus improves search effectiveness (Vakkari, 2002; Sihvonen & Vakkari, 2004). And searcher's prior experience about the search topic is related to the perceived usefulness of additional search terms from a thesaurus-enhanced search system (Shiri & Revie, 2006). These findings suggest that expanded thesaurus terms may be of particular benefit to domain expert searchers and those who had not searched the particular topic.

To gain a better understanding of searcher's perception of the search process and its

relation to search performance, we hypothesize:

H4. Searcher's perceptions of MeSH terms usefulness will agree with their search performance.

H4a. Searcher's perception of whether MeSH helped them will agree with their search performance in terms of precision and recall measures.

H4b. Searcher's perception of whether MeSH helped them will agree with their search performance in terms of time spent.

Overall, this study attempts to assess the impact of MeSH terms on search performance in an interactive search environment. We specifically consider the factors of user characteristics, exemplified by user's domain knowledge and search experience, and how these factors affect the search task of query term formulation when searchers are interacting with different kinds of IR systems, with and without the MeSH terms.

1.3 Overview of the Study

This study employed user-oriented evaluation methods for information retrieval systems. The user-oriented IR experiment was designed to answer the question: To what extent do MeSH terms improve search effectiveness of different kinds of users? The general procedure included pre-search background questionnaire, training session, search tasks, post-search questionnaire and brief interview after all search tasks.

We observed four different kinds of information seekers using an experimental information retrieval system: (1) search novices; (2) domain experts; (3) search experts and (4) medical librarians. The information needs were a subset of the relatively difficult topics originally created for the Text REtrieval Conference (TREC) Genomics Track 2004 ("TREC 2004 genomics track document set", 2005). Participants searched either

using a version of the system in which abstracts *and* MeSH terms (MeSH+) were displayed or another version in which they had to formulate their own terms (MeSH-). Effectiveness of retrieval was measured using the relevance judgments provided by TREC.

The results suggest that MeSH terms are more helpful in terms of precision for domain experts than for search experts. Users achieve the same level of search effectiveness regardless of whether MeSH terms were offered. In line with previous findings about the usefulness of controlled vocabulary and automatic indexing (e.g., Salton, 1968, 1972) in a laboratory operational environment, we demonstrate that automatic indexing techniques can be as competitive as controlled vocabulary within an interactive search environment.

More importantly, domain experts obtained the best search results and spent the most time using MeSH terms, but they did not perceive that MeSH terms were useful. Domain experts' perceptions about the usefulness of terms in abstracts were correlated with the precision measure. Since search novices issued the least queries in view of technical topics, they did not benefit from query reformulations. MeSH terms did not help search experts and medical librarians. These findings suggested that domain knowledge exceeds search training in searches on technical topics.

The dissertation is laid out in five chapters. The first chapter frames the issues and introduces research questions and hypotheses. Drawing from related literature in the areas of organization of information, information retrieval and information seeking and use, with particular reference to the biomedical domain, the second chapter provides an analytical view for the study. Chapter 3 and 4 detail the research methodology for

information retrieval system evaluation from the user's perspective and highlight study results. Chapter 5 discusses the role that user's domain knowledge plays in the perception of search tasks and search effectiveness, reflects on the evaluation issues in comparing the usefulness of manual and automatic indexing systems and concludes the dissertation by synthesizing the results and laying out future research directions.

CHAPTER 2 RELATED WORK

In this chapter, we review previous work related to the question of how useful search terms are for different kinds of users. We first discuss how to assess the usefulness of displayed index terms for information retrieval. Next, we discuss the design and methodology considerations that affect the research validity in the assessment of the human-developed index terms and automatic indexing techniques, with particular reference to the factors of retrieval system, search topic and user. To gain insights to the complexity of search tasks, we review work on query reformulation and user perception during the search process, as well as the impact of domain knowledge and search experience. We conclude that previous research results of comparing the search effectiveness of human-developed terms and automatic indexing techniques are inconclusive because it is difficult to separate the factors of systems, topics and users in an interactive search environment.

2.1 Assessment of Index Terms

The search effectiveness of human-developed vs. automatic index terms has frequently been raised in the information science literature. The questions associated with the usefulness of index terms are critical to the theory and practice of organizing information resources (see, e.g., Rowley, 1994; Svenonius, 1986; Dextre Clarke, 2008). Recent reports have suggested the use of automatic subject indexing because manual indexing is not cost effective (Calhoun, 2006) or difficult to understand and use (“Bibliographic Services Task Force,” 2005). The question of the usefulness of human-developed index terms is key to the operation of information retrieval systems in a

networked environment because library and information services are competing with other service providers, such as search engines.

Studies that directly compare the usefulness of manual and automatic indexing systems within laboratory controlled environments have suggested that automatic indexing methods, if implemented properly, can be as effective as manual indexing systems (see e.g., Salton, 1969, 1972). In a comparison of the Boolean search with human-developed controlled terms in MEDLARS (Medical Literature Analysis and Retrieval System) and the automatic vector matching techniques in SMART system, Salton (1972) claimed that “*fully automatic text processing methods can be used to obtain retrieval output of an effectiveness substantially equivalent to that provided by conventional, manual indexing* (emphasis original) (p. 81).” One of the limitations of the experiment is the scalability of retrieval techniques, with the use of relatively small test collection (450 bibliographic records in Salton, 1972; cf. Hersh, Buckley, Leone, & Hickam, 1994).

In Savoy’s (2005) evaluation of various search models using a relatively large test collection of 148,688 bibliographic records, the result reveals that the mean average precision obtained by the combined indexing strategy is significantly better than the single manual or automatic indexing schemes. Note that the distinction between manual and automatic indexing schemes is whether manually assigned index terms from the INIST (INstitut de l’Information Scientifique et Technique) thesaurus are included in the indexing and searching strategies. Since this information retrieval experiment was conducted in a laboratory environment without any human searchers, human-developed

index terms have not been properly assessed within the setting that is too removed from real life IR systems.

2.1.1 Assessment of human-developed index terms

Traditionally an index language refers to the language to describe the documents for retrieval purposes, whereas index terms are elements of the index language (Cleverdon, 1967; van Rijsbergen, 1979). One kind of human-developed index terms, such as thesauri and subject headings, is controlled in the sense that they are designed to address the problem of vocabulary variability by unifying term variants.

For example, a user searching on the term *bear*, meaning a large mammal was likely to retrieve documents on the *bear* stock market or on the right to *bear* arms. A user searching on the affective problem of *manic disorder* can use other words or phrases that mean nearly the same, such as *mania*, *bipolar disorder* and *bipolar depression*. The words or phrases that are chosen to represent these concepts can be used to form a controlled vocabulary.

A controlled vocabulary is designed to address the problem of “many-one and one-many relationships between words and their referents” (Svenonius, 1986, p. 332) in the use of natural language in information retrieval. In a sense, this device was intended to bridge the gap between the terms expressed by users with real information needs and the terms represented in the document collection. Since the work on controlled vocabulary was primarily based on the assumption that natural language is not systematic enough for representing and accessing information (see e.g., Svenonius, 1986, 2000), controlled indexing languages, such as Medical Subject Headings (MeSH) (“National Library of Medicine (U.S.)”, 1960), Thesaurus of Engineering and Scientific Terms

(TEST) (1967) and PREserved Context Index System (PRECIS) (Austin, 1976), were designed to minimize the ambiguities inherent in natural languages.

MeSH terms are recognized as one of the start-of-the-art controlled vocabularies for information retrieval systems (e.g., Aronson et al., 2000; Nelson, Johnston, & Humphreys, 2001; Humphrey et al., 2006), in part because they are continuously maintained by professional indexers in the National Library of Medicine (U.S.) and used to index the documents in MEDLINE database. MeSH terms are also used in PubMed, the public accessible version of MEDLINE on the Web. A search in PubMed on *bear*, meaning a large mammal will be automatically translated into the controlled MeSH term *ursidae* (technical name for a family of bears).

From a perspective of systems of vocabulary control, comparing the different characteristics of index languages will contribute to the development of manually identified term relations of important concepts in the document. However, some studies have suggested that complex term relations are not useful in terms of retrieval effectiveness. For instance, in a large-scale study comparing five types of index languages in the subject field of library and information science considering the factors of index language specificity and exhaustivity, and method of co-ordination, Keen (1973) shows that there were no large differences in retrieval effectiveness and efficiency for index languages. Sparck Jones' (1981) review of 1958-1978 index language tests suggests that different index languages can achieve comparable levels of search performance; more importantly, simple indexing is as good as sophisticated indexing. But these system evaluations were conducted in a laboratory environment.

Some empirical studies have suggested that users have difficulty understanding the meaning of controlled index languages when index terms occur in various subdivision orders and contexts of display (e.g., Drabenstott, Simcox, and Williams, 1999; Keen, 1977). To make the best use of human-developed index terms, it is important to determine whether the intended users of index terms are able to correctly interpret the intended meaning of these terms as assigned by human indexers, and assess what the impact of user's interpretations has on search effectiveness.

Studies from library online catalogs have demonstrated that the value of subject headings lies in the retrieval of relevant records not retrieved by keyword searching alone (e.g., Gross & Taylor, 2005; Voorbij, 1998). However, these studies implicitly assume that records not retrieved by keyword searching would be relevant to the query. More importantly, most users seemed to use keyword searching in part because they have difficulty formulating queries with subject headings (e.g., Larson, 1991). The question of the usefulness of human-developed index terms is key to the design of information retrieval systems partly because the development of human created controlled vocabularies is resources intensive compared with other automatic indexing techniques (see e.g., "Bibliographic Services Task Force," 2005; Calhoun, 2006).

More recently, Wacholder and Liu (2006, 2008) specifically compared the usefulness of query terms, traditionally called index languages, identified by different methods: one constructed by a human indexer and two others identified automatically. The prominent findings are that query languages affect search outcome and that a set of automatic terms using linguistically-motivated rules can be as effective as terms identified by a human indexer in an interactive search environment.

Overall, information search is a complex process. Despite the fact that index languages do not make large differences in search performance, previous research has revealed that the usefulness of human-developed index terms is affected by several factors, such as the subject domain of document collection, nature of controlled vocabulary and skills of indexers and searchers (Svenonius, 1986). A controlled user experiment for assessing the specific impact of particular variables on search performance is complex because it is relatively difficult to isolate these factors (Anderson & Pérez-Carballo, 2001). A study that is designed to assess the impact of subject domain, characteristics of index languages, or searcher characteristics on user search performance is challenging in itself because of the difficulty of separating the factors and the possible interactions among them.

2.1.2 Assessment of automatic indexing techniques

From a perspective of information retrieval, researchers are primarily interested in the effect of different automatic retrieval techniques on search performance in terms of the precision and recall measures. The choice of performance measures of precision and recall in Cleverdon's (1967) second Cranfield project has been widely used in evaluating the effectiveness of automatic indexing techniques. The precision measure, the ratio of retrieved relevant documents over the retrieved documents, is intended to assess how well an IR system can reject non-relevant documents; the recall measure, the ratio of retrieved relevant documents over the total relevant documents, is proposed to determine how well an IR system can obtain more relevant documents. This design and methodology has been very successful probably because researchers can test the performance of different retrieval techniques in a laboratory environment.

The test design and methodology following the Cranfield paradigm culminated in the TREC (Text REtrieval Conference) activities since the 1990s. TREC has provided a research forum for comparing the search effectiveness of different retrieval techniques across IR systems in a laboratory and controlled environment (Harman, 1993; Voorhees & Harman, 2005). The very large test collection used in TREC provided a test bed for researchers to experiment the scalability of retrieval techniques, which had not been possible in previous years.

Besides the standard ‘ad hoc’ search tasks, the tracks within the TREC Conference are designed to encourage new research areas; different tracks are performed each year as researchers’ interests change or more important questions are raised (Voorhees & Harman, 2005). The introduction of Interactive Track in 1992 demonstrates interests in user interaction with IR systems, whereas the Genomics Track initiated in 1993 focuses on the application of IR systems in a specific domain (see, Dumais & Belkin, 2005; Hersh, Bhuptiraju, Ross, Johnson, Cohen, & Kraemer, 2004). Although each track has specific areas of interest, the evaluation methodology principally applies to the test collection, search topics and relevance-judged document set.

The TREC search topics were constructed to simulate statements of user information needs and provide a clear description of what criteria that make document relevant. However, the artificiality and manipulation of topics have been one of the most frequently raised questions (see e.g., Beaulieu, Robertson & Rasmussen, 1996; Sparck Jones, 2005). But these topics are appropriate for TREC in the sense that different retrieval techniques can be compared with a baseline to enhance system performance.

Another important consideration is the variability of search topics in IR evaluation. Reflecting on the TREC Conferences, Sparck Jones (2000, p. 37) stated, “the nature and treatment of the user’s request is by far the dominant factor in performance.” Because of the potential interfering effect of search topic variability on search performance, researchers have used a very large number of search topics (e.g., fifty topics in TREC main search tasks; see Voorhees & Buckley, 2002 for the choice of topic size) and proposed a new typology for controlling search topics if available (Robertson, 1981) or a non-matched-pair experimental design where two different sets of topics are used in tests (Robertson, 1990).

In a controlled user IR experiment, within the constraint of human effort, it is not feasible ask participants to search a large number of topics within a single search session. Still, the result from a study that compared experimental interactive IR systems across sites in the TREC-6 interactive track using six topics reveals the dominating effect of search topics (Lagergren & Over, 1998). This suggests that putting human searchers in the loop of IR experiments introduces another source of variability that makes it difficult to distinguish the effect of systems, topics and users and their possible interactions (Voorhees, 2008).

To accurately assess the search performance in terms of the precision and recall measures, it is crucial to ensure the reliability of relevance judgment set. In order to assess how the inherently subjective relevance judgments may influence the measurement of retrieval effectiveness in TREC experiments, Voorhees (2000) verified the reliability of the relevance judgments in TREC. Relevance judgments were created using pooling methods from participating teams, and a single experienced assessor measured relevance

on a binary scale. The result shows that the relative reliability of relevance judgments using the pooling method and an experienced human assessor is appropriate for the improvement and comparison of IR techniques in a laboratory environment.

An overview of related work on the assessment of human-developed index terms and automatic indexing techniques reveals the complexity and importance of a good methodology for evaluating the usefulness of index terms in user information access. Despite the fact that different index languages do not make substantial differences in retrieving relevant documents in laboratory controlled experiments, it is crucial to develop well-designed studies that can separate the factors of systems, topics and users in an interactive search environment. To gain insights into search processes and user search behaviors, we will review related work on the search techniques designed to support query reformulations and user perceptions during the search process.

2.2 Query Reformulation and User Perception

In this section, we will review works on augmenting user queries in support of query reformulation tasks, specifically automatic query expansion and interactive query expansion, and user perceptions about the usefulness of expanded terms.

2.2.1 Query reformulation

Query reformulation, user's articulation of information needs after the initial search, has been considered an important component in information retrieval systems because users have problems articulating their information needs. It is also recognized that user queries are underspecified in the sense that they are typically very short representations of complex information needs.

Researchers have implemented several ways of augmenting user queries automatically by using indexing thesauri or dictionaries (e.g., Salton & Lesk, 1968; Srinivasan, 1996). Because the expanded terms are not displayed to human searchers for visual inspection in these systems, most studies in this area have focused on the usefulness of various techniques of term augmentation in retrieval effectiveness-based laboratory evaluations.

Several automatic query expansion studies that used the MEDLINE collection and associated controlled vocabulary MeSH terms have shown that properly implemented search systems with automatic query expansion can improve search performance, but the magnitude of improvement depends on the treatment of user queries, source and identification method of expanded terms and retrieval models (see e.g., Abdou & Savoy, 2008; Ijzereef, Kamps, & de Rijke, 2005; Lu, Kim and Wilbur, 2009). More importantly, since these studies do not specifically consider user behavior of query formulation or reformulations, the improvement in system performance in laboratory settings may not translate into user search performance in realistic situations.

Another line of research concerning user interactions with automatically suggested terms, namely interactive query expansion, has investigated users' selection of expanded terms and their perception of the semantic relationships of those terms. For example, a study of term selection behavior using the INSPEC database showed that 66% of the term relationships between original query terms and the best five terms users selected from the ranked list, are hierarchically related (Efthimiadis, 2000). In a study that examined user term selection behavior of expansion terms that were automatically extracted from the top 25 retrieved documents, Ruthven (2003) found that users were not

always able to identify the semantic relationships between query terms and expansion terms. In particular, users were not always able to identify semantic relationships useful for retrieving more relevant documents. The results from Joho, Sanderson and Beaulieu (2004) suggested that users were not consciously aware of the difference between two lists of terms with hierarchies, even though accessing the hierarchies reduced user efforts and increased the chance of finding relevant documents than the baseline system.

More recently, some researchers have been concerned with the assessment of displayed index terms within an interactive information access system. For example, Wacholder and Liu (2006, 2008) specifically compared the usefulness of query terms, traditionally called index languages, for question-answering tasks in a book. Because the design of search interface forced users to directly access the text by a list of displayed index terms, this study was able to compare user preference for these terms.

Overall, these studies suggest that properly implemented automatic query expansion devices in general improved system performance, but the improvement partly depended on the treatment of user queries. For interactive query expansion, users without specialized training in database searching or with limited knowledge about the collection were unable to identify useful terms for retrieving more relevant documents. To better support query reformulation tasks, we need to specifically consider the user characteristics, treatment of user queries, source and identification method of expanded or displayed terms, retrieval models and search interfaces. We will focus on user perceptions of the usefulness of displayed terms during the search process in the next session.

2.2.2 User perceptions of the usefulness of displayed terms

User perceptions of the search process are an important aspect of query reformulation tasks in an interactive search environment. To make search tools like displayed index terms useful for complex query reformulation tasks, one of the fundamental questions is what influences user perceptions of the usefulness of displayed terms, and whether users are able to recognize potentially useful terms in view of search topics and ultimately obtain better search results.

In a large-scale study of search behavior of professional online searchers, Fidel (1991) showed that the perceived quality of index terms affected the use of descriptors in online search activities. In a study of end-users' perceptions of a thesaurus-enhanced search interface, Shiri and Revie (2005) indicated that users with varying levels of domain knowledge had different perceptions about the potential benefit of using thesauri; more importantly, searcher's prior experience about the search topic was correlated with the perceived usefulness of additional suggested terms (Shiri & Revie, 2006). Studies of interactive query expansion showed that domain experts' selection of expanded terms from a thesaurus improved search effectiveness (Vakkari, 2002; Sihvonen & Vakkari, 2004).

Overall, these studies suggest that the relationship between the perceived usefulness of displayed terms and user search performance may depend on what kinds of users they are in an interactive search environment. In the next session, we will look at the impact of user characteristics of domain knowledge and search experience.

2.3 User Characteristics

Research in interactive information seeking has demonstrated the impact of user differences, but none of the studies mentioned above explicitly took such differences into account. In particular, domain knowledge and search experience have been identified as important research variables in the investigation of user search behaviors since the beginning of online intermediary searching during the 1970s (see e.g., Barry et al., 2005; Moore, Erdelez, & Wu, 2007; Wildemuth, 2004).

Studies in user search behaviors have generally suggested that there are large individual differences in search performance, even within a user group distinguished by different levels of either domain knowledge or search experience. From a perspective of IR system evaluation, one common limitation of these studies is that a relatively small number of search tasks (or named search topics in TREC) has been used without specific considerations of variations of these topics that may be responsible to the insignificant differences in aggregated search results (see e.g., Sparck Jones & van Rijsbergen, 1976; Buckley & Voorhees, 2005). That is, the impact of user characteristics on search results has not been properly assessed due to the dominating effect of search topics.

Despite these limitations, some empirical studies have indicated some results that may be run counter to the researcher's intuitions:

- (a) Domain knowledge or specific topic knowledge is not correlated to search outcome (Allen, 1991; Pao et al., 1993);
- (b) Search experience with databases cannot predict search outcome (Fenichel, 1981; Howard, 1982; Sutcliffe et al., 2000); and

- (c) There is an interaction effect between domain knowledge and search experience (Hsieh-Yee, 1993; Meadow, Marchionini, & Cherry, 1994; Vakkari, Pennanen, & Serola, 2003).

2.3.1 Domain knowledge

Domain knowledge refers to an individual's level of knowledge in a particular subject discipline. This variable has been operationalized and measured in several ways, depending on the purposes of different studies. For example, medical students' clinical knowledge was measured by standardized tests, the University of Michigan's Comprehensive Clinical Assessment examination and Part II of NBME (National Board of Medical Examiners), since the study population was medical students (Pao et al., 1993). Hembrooke and her colleagues' (2005) study used the subject's self-report of search topic familiarity as a measure of the domain expertise since the study was designed to investigate undergraduate students' Web searching behavior. To study the effect of the domain knowledge on user search behavior, stages of semester long course instruction in a particular field (Sihvonen & Vakkari, 2004) and formal training in a subject domain (Hsieh-Yee, 1993; Marchionini & Dwiggins, 1990; Meadow, Wang, & Yuan, 1995; Wildemuth, 2004) were also used as a measure of the participant's level of knowledge.

Surprisingly, studies that investigated the effect of domain knowledge on the search effectiveness in an information retrieval system have shown that these two variables are not correlated. For example, in a study of medical students' use of MEDLINE, Pao and her colleagues (1993) found that there is no relationship between search effectiveness and medical students' clinical knowledge measured by standardized

medical tests. Allen's (1991) study of students' use of an online catalog showed that there is no correlation between the specific topic knowledge and the approximate recall of search results. The main difference is that high topic knowledge users used more search expressions than low topic knowledge users. As noted earlier, both studies have not specifically considered search topic variability and used a very small number of search topics.

However, users' level of domain knowledge has been shown to affect the use of thesaurus terms implemented on a search system. Some studies have suggested that users who have reasonable understanding about the search topic are able to make use of thesaurus tools in the retrieval task of query formulations. For instance, users in academic environment found it informative and useful to use the INSPEC thesaurus navigation feature for query enhancement (Jones, Gatford, Robertson, Hancock-Beaulieu, Secker, & Walker, 1995). In a test of whether an enhanced thesaurus will be useful in a real work environment, Nielsen (2004) showed that domain experts perceived that a word association thesaurus was useful for query formulation. None of these studies, however, has established the relationship between user perception of displayed terms' usefulness and search effectiveness, or distinguished whether domain experts will benefit more from the use of thesaurus terms than domain novices.

2.3.2 Search experience

Search experience refers to searcher's skills in interacting with an information retrieval systems. For studies that were conducted in the 1980s and early 1990s, end-users usually had limited experiences searching online bibliographic databases because online searching was very expensive and professional librarians usually conducted the

search on behalf of users. Here search experience usually referred to whether searchers have had extensive use of online databases and whether they were proficient in the system features, such as search commands or indexing thesauri.

For example, the search experience was measured by the total number of searching sessions in a longitudinal study of medical students' use of MEDLINE (Pao et al., 1993). Several studies that examine the effect of search experience on searching behavior have used the total time spent using a particular online database or Dialog system as a measure of different levels of search experience (Fenichel, 1981; Howard, 1982; Yuan, 1997). Other studies that investigated whether search success depends on searchers' personal characteristics, the search experience was determined by formal training in online database searching (Bellardo, 1985; Saracevic & Kantor, 1988).

More recent studies tend to assess whether the search experience in a specific type of information retrieval system can be transferred to another. For example, since one of the primary objectives was to investigate the effect of online database search experience on Web search performance, Palmquist and Kim (2000) used the duration and frequency of using online databases to measure undergraduate students' search experience. Because of the similar system features in Boolean logic, Vakkari, Pennanen and Serola (2003) used the frequency of online public access catalog as a measure of undergraduate students' search experience in a Boolean-based online database. However, it is still not known whether the search experience is transferrable within the Boolean-based IR system or between different types of IR systems.

Despite different measurement in above-mentioned studies, the study of the impact of search experience on search performance will provide a rationale for formal online search training.

In summary, our review of the research literature reveals the complexity and importance of a good methodology for evaluating the usefulness of index terms in user information access. Earlier studies (e.g., Cleverdon, 1967; Keen, 1973) evaluated the usefulness of index terms in a laboratory environment. Previous research results of comparing the search effectiveness of human-developed terms and automatic indexing techniques in an interactive search environment are inconclusive because of the difficulty of separating out factors of systems, topics and users. The relationship between the perceived usefulness of index terms and user search performance may depend on different kinds of users, distinguished by levels of domain knowledge and search experience.

The next chapter features our approach to the overall question of the usefulness of MeSH terms for different kinds of users. It describes in detail the experimental design and the rationale for this design.

CHAPTER 3 RESEARCH METHODOLOGY

This chapter describes the research design we used to assess the usefulness of MeSH terms for different kinds of users. We employed a user-oriented evaluation methodology to assess search effectiveness of automatic and manual indexing methods in an interactive information retrieval environment. Our approach used a relatively large number of search topics, careful attention to experimental design and a complex Greco-Latin square design.

3.1 Overview of Method

Thirty-two searchers from a major public university and nearby medical libraries in the northeast area of the US participated in the study. Each searcher belonged to one of four groups:

- Search Novice (SN)
- Domain Experts (DE)
- Search Experts (SE)
- Medical Librarians (ML)

The experimental task was to conduct a total of eight searches to help biologists conduct their research. Participants searched either using a version of the system in which abstracts *and* MeSH terms were displayed (MeSH+) or another version in which they had to formulate their own terms based only on the display of abstracts (MeSH-). Participants conducted four searches each with two different systems: in one, they browsed a displayed list of MeSH terms (MeSH+) and in the other (MeSH-). Half the participants used MeSH+ system first; half used MeSH- first.

Search topics were selected from the topics used in TREC 2004; these topics, which were developed for automatic searching, are relatively difficult (see Appendix F). We also used the relevance judgments originally created for measuring the search effectiveness of information retrieval techniques. The document set consisted of 3,442,321 bibliographic records with abstracts from the 2004 TREC (Text REtrieval Conference) Genomics document set (“TREC 2004 genomics track document set,” 2005). We decided to use these difficult search topics because of the availability of the TREC relevance judgments, and because it allows us to compare the usefulness of human created terms to standard retrieval techniques.

To help motivate participants to do their best, we promised monetary incentives according to their search performance. We were concerned that difficult tasks may prevent participants from completing all searches, and that the motivational characteristics of participants are possible sources of sample bias (Sharp, Pelletier & Levesque, 2006). The experimental setting for most searchers was a university office; for some searchers, it was a medical library. Before they began searching participants were briefly trained in how to use the MeSH terms. We kept search logs that recorded search terms, a ranked list of retrieved documents, and time-stamps.

3.1.1 Subjects

We used the purposive sampling method for recruiting our subjects since we were concerned with the impact of specific searcher characteristics on search effectiveness. The key searcher characteristics were different levels of domain knowledge in the biomedical domain and whether they had substantial search training. The four types of

searchers were distinguished by their levels of domain knowledge and search training (see Figure 3-1).

	Domain Knowledge	Search Training
Search Novice (SN)	–	–
Domain Experts (DE)	+	–
Search Experts (SE)	–	+
Medical Librarians (ML)	+	+

Note. Plus (+) and minus (–) indicate the high-level and low-level of the specified searcher characteristics respectively.

Figure 3-1 Four types of searchers categorized by domain knowledge and search training

The four kinds of searchers were operationalized as follows:

1. Search Novices (SN). Undergraduate students without formal training in online searching courses and without advanced knowledge in biomedical domain. They are undergraduate students who are not biology majors. While many of these students are experienced and heavy Web users, they are not expected to have in-depth understanding about online bibliographic databases.
2. Domain Experts (DE). Graduate students in a biomedical domain, i.e., biology or medicine. DEs did not have formal training in searching, such as online searching courses.
3. Search Experts (SE). Graduate students enrolled in Master of Library and Information Science (MLIS) programs who had previously taken online database searching or other related courses and do not have advanced knowledge in

biomedical domain. SEs had not majored in biology and did not have a Master degree or above in any biomedical field.

4. Medical Librarians (ML). Medical librarians specializing in online searching services. The domain knowledge is defined by formal education in biomedical areas or more than two-year experience of working in medical libraries.

A total of thirty-two searchers (8 for each type of searchers) participated in the study. They were assigned to one of four categories based on different sources of contact and two initial questions: (1) Have you taken any college-level biology courses? (2) Have you taken classes in how to do online searching? This selection was made to ensure that all searchers are representative of each category. All participants had extensive Web search experience. More than two-thirds reported that they use search engines every day or several times a day or more.

3.1.2 Experimental design

The experiment was a $4 \times 2 \times 2$ factorial design with four types of searchers, two versions of an experimental system (MeSH+ and MeSH-) and controlled search topic pairs. The versions of a system, types of searchers (distinguished by levels of domain knowledge and search training) and search topic pairs were controlled by a Graeco-Latin square balanced design (Fisher, 1935). The possible ordering effects have been taken into account by the design. The requirement for this experimental design is that the examined variables do not interact and each variable has the same number of levels (Kirk, 1995). The treatment layout of a 4×4 Graeco-Latin square design is illustrated in Figure 3-2.

1	2	3	4	5	6	7	8
SN	DE	SE	ML	DE	SN	ML	SE
38	12	29	50	38	12	27	45
12	38	50	29	12	45	38	27
29	50	12	38	27	38	45	12
50	29	38	12	45	27	12	38
42	46	32	15	9	36	30	20
46	42	15	42	36	9	20	30
32	15	42	46	30	20	9	36
15	32	46	32	20	30	36	9
9	10	11	12	13	14	15	16
SE	ML	SN	DE	ML	SE	DE	SN
29	50	27	45	42	46	9	36
50	29	29	27	46	36	42	9
27	45	45	50	9	42	36	46
45	27	50	29	36	9	46	42
2	43	1	49	2	43	33	23
43	1	49	2	43	2	23	33
1	49	2	43	33	23	2	43
49	2	43	1	23	33	43	2

Numbers 1-16 refers to participant ID; SN, DE, DE and ML refer to types of searchers, SN=Search Novices, DE=Domain Experts; SE=Search Experts; ML=Medical Librarians; Red and blue color blocks refer to MeSH and Non-MeSH versions of an experimental system; Numbers in color blocks refer to search topic ID number from TREC Genomics Track 2004 data set; Ten search topic pairs, randomly selected from a pool of twenty selected topics, include (38, 12), (29, 50), (42, 46), (32, 15), (27, 45), (9, 36), (30, 20), (2, 43), (1, 49) and (33, 23).

Figure 3-2 4×4 Graeco-Latin square design

Because of the potential interfering effect of search topic variability on search performance in IR evaluation, we used a design that included relatively large number of search topics. In theory, the effect of topic variability and topic-system interaction on system performance could be eliminated by averaging the performance scores of the topics (micro-averaging method), together with the use of very large number of search topics (e.g., fifty topics in TREC main track evaluation activities). The TREC standard ad hoc task evaluation studies (Banks, Over & Zhang, 1999; Buckley & Voorhees, 2005) and other proposals of ideal test collections (e.g., Robertson, 1981, 1990; Sparck Jones & van Rijsbergen, 1976) have been concerned with the large variability in search topic performance. However, in a user-centered IR experiment it is not feasible to use as many as fifty search topics because of human fatigue.

In this study we controlled search topic pairs by a balanced design in order to alleviate the overriding effect of search topic variability. We assumed that all the search topics are equally difficult, since we do not have a good theory about what makes some search topics more difficult than others. By design we ensured that each search topic pair was assigned to all types of searchers and was searched at least two times by the same type of searchers. This design required a total of ten search topic pairs and a minimum of sixteen participants.

3.1.3 Search tasks and incentive system

The search task was designed to simulate online searching situations in which professional searchers look for information on behalf of users. We decided to use this relatively challenging task for untrained searchers because choosing realistic tasks such as this one would enhance the external validity of the experiment. Considering the relatively difficult tasks, searchers may have problems completing all searches. And because research literature has suggested that the motivational characteristics of participants are possible sources of sample bias (Sharp, Pelletier & Levesque, 2006), we designed an incentive system to motivate the searchers.

We promised monetary incentives according to the participant's search effectiveness. Each subject was paid \$20 for participating and was also paid up to \$10.00 dollars more based on the average number of relevant documents in the top ten search results across all search topics; on average each participant received an additional \$4.40, with a range of \$2.00 - \$8.00.

3.1.4 Experimental procedures

An overview of experimental procedures is provided in Figure 3-3. After signing the consent form, the participant filled out a searcher background questionnaire before the search assignment (see Appendix A and B for informed consent form and searcher background questionnaire). After a brief training session, they were assigned to one of the arranged experimental conditions and conducted search tasks. They completed a search perception questionnaire and were asked to indicate the relevance of two pre-judged documents when there were done with each search topic (see Appendix C for post-search questionnaire). A brief interview was conducted when they finished all search topics (see Appendix D for interview questions). Search logs with search terms and ranked retrieved documents were recorded.

Background Questionnaire	Training Session	Concept Analysis	Experimental System	Post-Search Questionnaire	Follow-up Interview
		MeSH Browser	MeSH+ Version		
		Come up with Terms	MeSH- Version		

Figure 3-3 An overview of experimental procedures

To ensure that the participant received consistent training, an experimental guideline with scripted instructions in colloquial English, together with a training search topic, was prepared and used. This training session was designed to familiarize participants with available system features and search tasks, particularly search concepts formulations and examination of search results. A sample document record was used to illustrate the availability of particular index terms in which MeSH terms were only accessible half of the time. A search help with advanced system features was provided to

the participant on a piece of paper as part of the tutorial (see Appendix E for experimental guidelines, sample document record and search help).

To help searchers recognize potentially useful search terms, participants were then instructed to do concept analysis by identifying important concepts from search topic descriptions and devising other terms within each concept (Figure 3-4). The chosen practice topic consisted of 3 (hypertension, genetic risk and stroke) or 4 (hypertension, risk factors, genetics and stroke) main concepts in order to illustrate different ways of analyzing search topics. The same concept analysis form printed with assigned topics was used and collected. We instructed participants to consult the MeSH Browser for coming up with other terms. All types of searchers seemed to understand this process, although only trained searchers had received this kind of training before the experiment.

Concept Analysis Form

Search Topic

ID: 39

Title: Hypertension

Need: Identify genes as potential genetic risk factors candidates for causing hypertension.

Context: A relevant document is one which discusses genes that could be considered as candidates to test in a randomized controlled trial which studies the genetic risk factors for stroke.

	Concept 1	Concept 2	Concept 3	Concept 4
Terms from description	hypertension	genetic risk	stroke	
		risk factors	genetics	stroke

Other terms	high blood pressure	relative risk	cerebrovascular accident	

Figure 3-4 Concept analysis form for training session

The MeSH Browser, an online vocabulary look-up aid, prepared by U.S. National Library of Medicine, was designed to help searchers find appropriate MeSH terms and display hierarchy of terms for retrieval purposes (see Figure 3-5 for a screenshot of MeSH Browser and hierarchical display of the MeSH term hypertension). The 2003 MeSH Browser was used to align with the test collection (“MeSH Browser (2003 MeSH),” 2004; “TREC 2004 genomics track document set,” 2005). The MeSH Browser was only available when participants were assigned to the MeSH+ version of an experimental system; in the MeSH– version, participants had to formulate their own terms without the assistance of MeSH Browser and displayed MeSH terms in bibliographic records.

MeSH Browser

http://www.nlm.nih.gov/mesh/2003/MBrowser.html

NATIONAL LIBRARY OF MEDICINE **MEDICAL SUBJECT HEADINGS** **MeSH**

[MeSH Home](#) | [Contact NLM](#) | [Site Index](#) | [Search Our Web Site](#) | [NLM Home](#)

[Health Information](#) | [Library Services](#) | [Research Programs](#) | [New & Noteworthy](#) | [General Information](#)

MeSH Browser (2003 MeSH):
The files are updated every week on Sunday.
[Go to 2004 MeSH](#)


Enter term or the beginning of any root fragments: OR

Search for these record types:

☐ Main Headings
☐ Qualifiers
☐ Supplementary Concepts
☒ All of the Above
☐ Search as MeSH Unique ID
☐ Search as text words in Annotation & Scope Note

☐ Search in these fields of chemicals:

☐ Heading Mapped To (HM) (Supplementary List)
☐ Indexing Information (II) (Supplementary List)
☐ Pharmacological Action (PA)
☐ CAS Registry/EC Number (RN)
☐ Related CAS Registry Number (RR)

 [MeSH vocabulary suggestions](#)

[About MeSH Browser](#) | [MeSH Home Page](#) | [Questions or Comments](#)
[NLM Classification, the scheme used to categorize and organize books, audiovisuals, and similar materials.](#)

[U.S. National Library of Medicine](#), 8600 Rockville Pike, Bethesda, MD 20894
[National Institutes of Health](#)
[Department of Health & Human Services](#)
[Copyright and Privacy Policy](#)
 Last updated: 16 August 2004

Done

Hypertension, Malignant

http://www.nlm.nih.gov/cgi/mesh/2003/M8.cgi?index=6769

National Library of Medicine - Medical Subject Headings

2003 MeSH

MeSH Descriptor Data

[Return to Entry Page](#)

MeSH Heading	Hypertension, Malignant
Tree Number	C14.907.489.330
Annotation	malignant does not refer to neoplasm: refers to severe hypertension with papilledema & arterial necrosis
Scope Note	Severe hypertension characterized by papilledema and necrosis of small arteries and arterioles. The diastolic pressure is generally greater than 130 mm Hg.
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH UR US VE VI
Unique ID	D006974

MeSH Tree Structures

[Cardiovascular Diseases \[C14\]](#)

[Vascular Diseases \[C14.907\]](#)

[Hypertension \[C14.907.489\]](#)

 ▶ [Hypertension, Malignant \[C14.907.489.330\]](#)

[Hypertensive Encephalopathy \[C14.907.489.330.500\]](#)

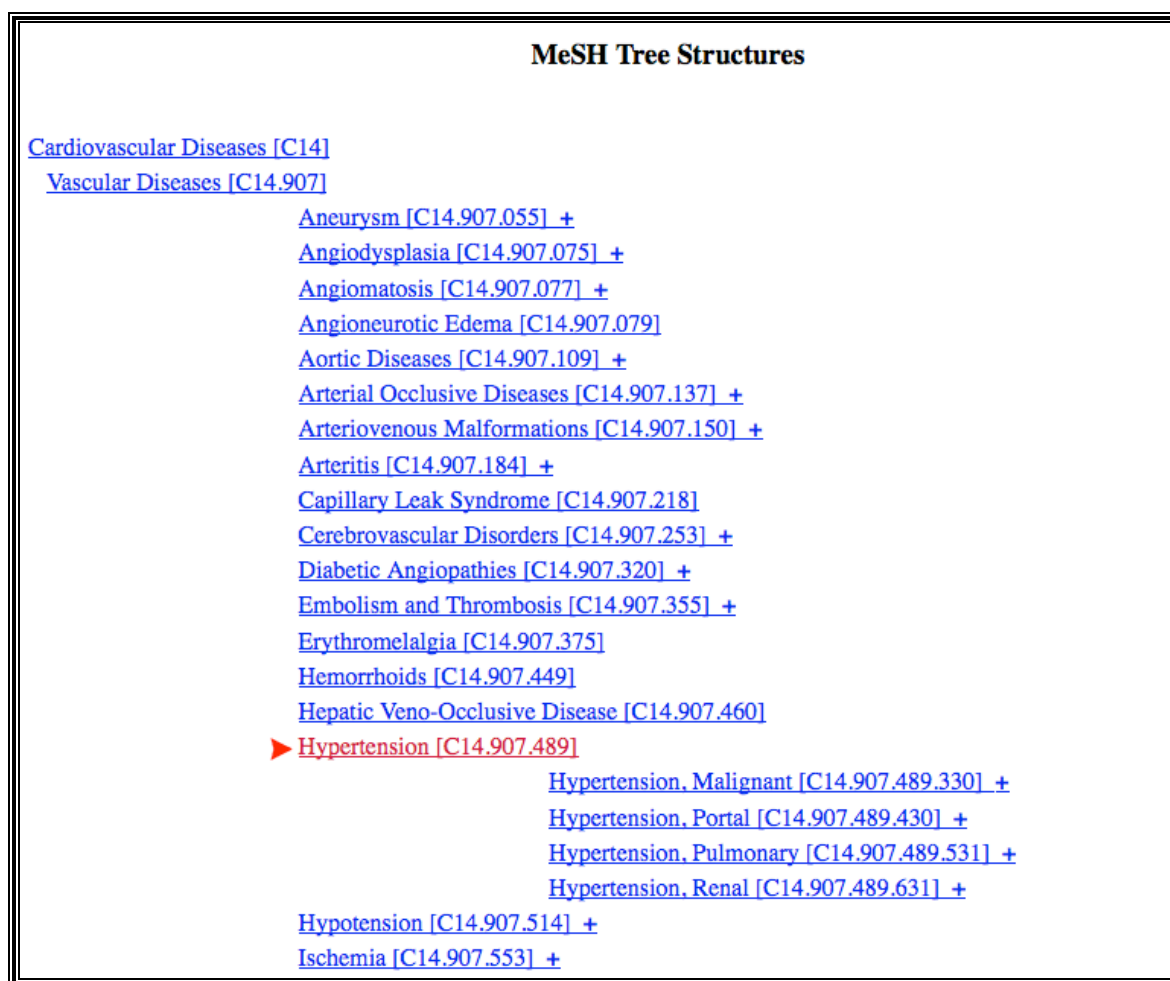
[Hypertension, Portal \[C14.907.489.430\]](#) +

[Hypertension, Pulmonary \[C14.907.489.531\]](#) +

[Hypertension, Renal \[C14.907.489.631\]](#) +

[Return to Entry Page](#) [Link to NLM Cataloging Classification](#)

Done



Source: “MeSH Browser (2003 MeSH)” (2004). Retrieved February 14, 2009, from <http://www.nlm.nih.gov/mesh/2003/>

Figure 3-5 The 2003 MeSH Browser and hierarchical display of the MeSH term hypertension

Participants were told that several biologists have certain information needs, which were described in the specific search topics. The task was to use the system to find as many relevant documents as possible. Each participant searched eight topics in total (four for each version of the system) from a pool of ten selected search topic pairs. Because we were concerned that the topics were so hard that even the medical librarians would not understand them, we used a questionnaire regarding search topic

understanding after each topic. The testing items of two randomly selected pre-judged documents, one definitely relevant and the other definitely not relevant, were prepared from the data set (“TREC 2004 genomics track document set,” 2005).

Each search topic was allocated up to ten minutes. The last search within the time limit was used for calculating search performance. To keep the participants motivated and reward their effort, they were asked to orally indicate which previous search result would be the best answer when the search task was not finished within ten minutes.

3.1.5 Experimental system

For this study, it was important for participants to conduct their searches in a carefully controlled environment; our goal was to offer as much help as possible while still making sure that the help and search functions did not interfere with our ability to measure the impact of the MeSH terms. We built an information retrieval system based on the Greenstone Digital Library Software version 2.70 (“New Zealand Digital Library Project,” 2006) because it provides reliable search functionality, customizable search interface and good documentation (Witten & Bainbridge, 2007).

We prepared two different search interfaces using a single system using Greenstone: MeSH+ and MeSH– versions (Figure 3-4). One interface allowed users to use MeSH terms; the other required them to devise their own terms. One interface displayed MeSH terms in retrieved bibliographic records and the other did not. Because we were concerned that the participant responds to the cue that may signal the experimenter’s intent, the search interfaces were termed ‘System Version A’ and ‘System Version B’ for ‘MeSH+ Version’ and ‘MeSH– Version’ respectively (see Figure 3-7 and Figure 3-8). The MeSH– version was used as baseline system for an automatic indexing

system, whereas the MeSH+ version served as performance of a manual indexing system.

That is, MeSH terms added another layer of document representation to the MeSH+ version.

	MeSH+ Version	MeSH- Version
MeSH Terms	+	-
Abstract	+	+

Note. The pluses and minuses refer to the existent and non-existent of the specified search fields on search interface and displayed terms in bibliographic records

Figure 3-6 Distinction between MeSH+ and MeSH- search interfaces

HOME HELP PREFERENCES

search

search titles a-z authors a-z dates phrases

Search and display results in order

Word or phrase (fold, stem) ... in field

and ☐ ☐ Full Records

and ☐ ☐ Title

and ☐ ☐ Abstract

MeSH Terms

Clear Form Begin Search

Or enter a query directly:

Run Query

search history

[food safety]:TX 1150+ results
(informal, ranked, case must match, stemmed,)

[canine]:TX 1150+ results
(ranked, case must match, stemmed,)

	search	titles a-z	authors a-z	dates	phrases
Authored By:	Katsuta Y; Zhang XJ; Aramaki T;				
Paper Title:	[PMID-11411134] [Pulmonary hypertension complicating portal hypertension: portopulmonary hypertension]				
Source:	Nippon Rinsho 2001 Jun;59(6):1186-92.				
Publication Date:	2001				
Abstract:	<p>Portopulmonary hypertension is a condition with a poor prognosis, which is defined as precapillary pulmonary hypertension complicating portal hypertension mainly due to cirrhosis of various etiologies. A mean pulmonary arterial pressure greater than 25 mmHg at rest with a pulmonary capillary wedge pressure less than 15 mmHg and a pulmonary vascular resistance greater than 120 dynes.sec.cm-5, in the setting of the presence of portosystemic shunting has been proposed as hemodynamic criteria for portopulmonary hypertension. Prevalence of pulmonary hypertension ascertained by right cardiac catheterization was 2% among patients with cirrhosis, and reached to 4% particularly among candidates for liver transplantation. Hyperdynamic systemic circulation seen commonly in patients with cirrhosis appeared to be normalized by complication of pulmonary hypertension with a contraction of circulating plasma volume. Long term treatment by epoprostenol administration or nitric oxide inhalation could induce a gradual decline in pulmonary arterial pressure in patients with poor response to acute vasodilator administration.</p>				
MeSH Terms:	[MeSH terms] Adolescent; Adult; Aged; English Abstract; Female; Hepatitis, Chronic/complications; Human; Hypertension, Portal/*complications/drug therapy; Hypertension, Pulmonary/drug therapy/*etiology; Liver Cirrhosis/*complications/physiopathology; Male; Middle Aged; Pulmonary Circulation; Vasodilator Agents/therapeutic use;				

Figure 3-7 MeSH+ version search interface and search output

[HOME](#) [HELP](#) [PREFERENCES](#)

search

search titles a-z authors a-z dates phrases

Search and display results in ranked order

Word or phrase (fold, stem) ... in field

<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>	Full records
and <input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>	Title
and <input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>	Abstract
and <input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>	Author

Or enter a query directly:

search history

☐ [hypertension]:TI 1150+ results
(informat, ranked,
case must match,
stemmed,)

☐ [food safety]:TX 1150+ results
(ranked, case must
match, stemmed,)

search		titles a-z	authors a-z	dates	phrases
Authored By:	Katsuta Y; Zhang XJ; Aramaki T;				
Paper Title:	[PMID-11411134] [Pulmonary hypertension complicating portal hypertension: portopulmonary hypertension]				
Source:	Nippon Rinsho 2001 Jun;59(6):1186-92.				
Publication Date:	2001				
Abstract:	<p>Portopulmonary hypertension is a condition with a poor prognosis, which is defined as precapillary pulmonary hypertension complicating portal hypertension mainly due to cirrhosis of various etiologies. A mean pulmonary arterial pressure greater than 25 mmHg at rest with a pulmonary capillary wedge pressure less than 15 mmHg and a pulmonary vascular resistance greater than 120 dynes.sec.cm⁻⁵, in the setting of the presence of portosystemic shunting has been proposed as hemodynamic criteria for portopulmonary hypertension. Prevalence of pulmonary hypertension ascertained by right cardiac catheterization was 2% among patients with cirrhosis, and reached to 4% particularly among candidates for liver transplantation. Hyperdynamic systemic circulation seen commonly in patients with cirrhosis appeared to be normalized by complication of pulmonary hypertension with a contraction of circulating plasma volume. Long term treatment by epoprostenol administration or nitric oxide inhalation could induce a gradual decline in pulmonary arterial pressure in patients with poor response to acute vasodilator administration.</p>				

Figure 3-8 MeSH– version search interface and search output

The experimental system was constructed as Boolean-based system with ranked functions by the TF×IDF weighting rule (Witten, Moffat, & Bell, 1999). More specifically, MGPP (MG++), a re-implementation of the mg (Managing Gigabytes) searching and compression algorithms, was used as indexing and querying indexer. Basic system features, including fielded searching, phrase searching, Boolean operators, case sensitivity, stemming and display of search history, were sufficient to fulfill the search tasks. The display of search history was necessary because it provided useful feedback regarding the magnitude of retrieved documents for difficult search tasks that usually required query reformulations.

Since our goal was specifically to investigate the usefulness of displayed MeSH terms, we deliberately refrained from implementing certain system features that allow users to take advantage of the hierarchical structures of MeSH terms, such as the hyperlinked MeSH terms, explode function that automatically includes all narrower terms

and automatic query expansion (see e.g., Hersh, 2008; Lu, Kim & Wilbur, 2009) available on other online search systems. The use of those features would have invalidated the results by introducing other variables at the levels of search interface and query processing, although a full integration of those system features would have increased the usefulness of MeSH terms.

Given that the participants possess varying levels of search skills, the search interface included search functions designed for each type of searchers. To make sure that trained searchers could use their advanced searching skills, they were able to directly construct complex Boolean queries in a large query box. But it would not have been a fair test of untrained searchers if they had to use Boolean queries, so we gave them an easier interface, a menu of Boolean options with four search fields; advanced searchers also had the option to use the more basic search interface.

3.1.6 Documents

The experimental system was set up on a server, using bibliographic records from the 2004 TREC Genomics document set (“TREC 2004 genomics track document set,” 2005). TREC Genomics Track 2004 Data Set document test collection was a 10-year (from 1994 to 2003) subset of MEDLINE with a total of 4,591,108 records. The test collection subset fed into the system used 75.0% of the whole collection, a total of 3,442,321 records, excluding the records without MeSH terms or abstracts.

We prepared two sets of documents for setting up the experimental system: MeSH+ and MeSH– versions. One interface allowed users to use MeSH terms; the other did not provide this search option. The difference was also reflected in retrieved bibliographic records.

3.1.7 Search topics

The search topics used in this study were originally created for TREC Genomics Track 2004 for the purpose of evaluating the search effectiveness of different retrieval techniques (see Figure 3-9 for an example). They covered a range of genomics topics typically asked by biomedical researchers. Besides a unique ID number for each topic, the topic was constructed in a format that included the title, need and context fields. The title field was a short query. The need field was a short description of the kind of material the biologists are interested in, whereas the context field provides background information for judging the relevance of documents. The need and context fields were designed to provide more possible search terms for system experimentation purposes.

ID: 39

Title: Hypertension

Need: Identify genes as potential genetic risk factors candidates for causing hypertension.

Context: A relevant document is one which discusses genes that could be considered as candidates to test in a randomized controlled trial which studies the genetic risk factors for stroke.

Figure 3-9 Sample search topic

Because of the technical nature of genomics topics, we considered whether the search topics are intelligible for human searchers, particularly for those without advanced training in the biomedical field. Given that these search topics were designed for machine runs with little or no consideration for searches by real users, we selected 20 of the 50 topics using the following procedure:

1. Consulting an experienced professional searcher with biology background and a graduate student in neuroscience, to help make a judgment as to whether the topics would be comprehensible to the participants who were not domain

experts. Topics that used advanced technical vocabulary, such as specific genes, pathways and mechanisms, were excluded;

2. Ensuring that major concepts in search topics could be mapped to MeSH by searching MeSH Browser. For instance, topic 39 could be mapped to MeSH preferred terms *hypertension* and *risk factors*;
3. Eliminating topics with very low MAP (mean average precision) and P10 (precision at top 10 documents) score in the relevance judgment set because these topics would be too difficult;
4. We selected a total of 20 search topics from a pool of 50 topics (see Appendix F for selected topics. These topics were then randomly selected to create ten search topic pairs for the arrangement of experimental conditions (see Figure 3-4).

3.1.8 Reliability of relevance judgment sets

We measured search outcome using standard precision and recall measures for accuracy and time spent for user effort (Cleverdon, 1967). Theoretically speaking, the calculation of recall measure requires relevance judgments from the whole test collection. However, it is almost impossible to obtain these judgments from a test collection with more than 3 million documents. For practical reasons the recall measure used a pooling method that created a set of unique documents from the top 75 documents submitted by 27 groups participated in the TREC 2004 Genomics Track ad hoc tasks (Hersh et al., 2004). Empirical evidence has shown that recall calculated with a pooling method provides a reasonable approximation, although the recall is likely to be overestimated (Zobel, 1998). But as a result of this approach, there was an average pool size of 976

documents, with a range of 476-1450, which had relevance judgments for each topic (Hersh et al., 2004).

It was quite likely that some of the participants in our experiment would retrieve documents that had not been judged. The existence of un-judged relevant documents, called sampling bias in pooling method, is concerned with the pool depth and the diversity of retrieval methods that may affect the reliability of relevance judgment set (Buckley et al., 2007). The assumption that the pooled judgment set is a reasonable approximation of complete relevance judgment set may become invalid when the test collection is very large.

To ensure that the TREC pooled relevance judgment set was sufficiently complete and valid for the current study, we analyzed top 10 retrieved documents from each human runs ($32 \text{ searchers} \times 8 \text{ topics} = 256 \text{ runs}$). Cross-tabulation results showed that about one-third of all documents retrieved in our study had not been judged in the TREC data set. More specifically, for a total of 2277 analyzed documents, 762 (33.5 %) had not been assigned relevant judgments. There existed large variations in percentage of un-judged documents for each search topic, with a range of 0–59.3% (Table 3-1).

Table 3-1 Distribution of TREC relevance judgments from top 10 documents retrieved by human participants

Topic	Documents with TREC relevance judgments		Documents with no associated TREC relevance judgment	Total
	Not Relevant	Relevant	Un-judged	
1	36	23	16	75
	48.0%	30.7%	21.3%	100.0%
2	64	13	80	157
	40.8%	8.3%	51.0%	100.0%
9	7	139	9	155
	4.5%	89.7%	5.8%	100.0%
12	20	63	56	139
	14.4%	45.3%	40.3%	100.0%
15	19	28	27	74
	25.7%	37.8%	36.5%	100.0%
20	39	19	18	76
	51.3%	25.0%	23.7%	100.0%
23	31	13	28	72
	43.1%	18.1%	38.9%	100.0%
27	73	59	28	160
	45.6%	36.9%	17.5%	100.0%
29	83	24	34	141
	58.9%	17.0%	24.1%	100.0%
30	15	51	3	69
	21.7%	73.9%	4.3%	100.0%
32	4	32	24	60
	6.7%	53.3%	40.0%	100.0%
33	18	9	20	47
	38.3%	19.2%	42.5%	100.0%
36	7	105	41	153
	4.6%	68.6%	26.8%	100.0%
38	34	50	74	158
	21.5%	31.7%	46.8%	100.0%
42	8	51	86	145
	5.5%	35.2%	59.3%	100.0%
43	4	119	20	143
	2.8%	83.2%	14.0%	100.0%
45	20	24	87	131
	15.3%	18.3%	66.4%	100.0%
46	14	82	22	118
	11.9%	69.5%	18.6%	100.0%

49	11	54	0	65
	16.9%	83.1%	0.0%	100.0%
50	27	23	89	139
	19.4%	16.6%	64.0%	100.0%
Total	534	981	762	2277
	23.5%	43.1%	33.5%	100.0%

To assess the impact of incomplete relevance judgments, we compared the top 10 ranked search results between the judged document set and the pooled document set for each topic. The judged document set was composed of the documents that matched TREC data, i.e., combination of judged not relevant and judged relevant in Table 3-1). The un-judged documents, added to the pooled document set, were considered ‘not relevant’ in our calculations of search outcome. The paired t-test results by search topic revealed significant differences between the two sets in terms of MAP ($t(19) = -3.69, p = .002, p < .01$), P10 ($t(19) = -3.89, p < .001$) and P100 ($t(19) = -3.95, p < .001$) measures. The mean of the differences for MAP, P10 and P100 was approximately 2.7%, 9.9% and 4.9% respectively. We concluded that the TREC relevance judgments were applicable to this study. In what follows, we calculated search effectiveness based on the relevance judgments in TREC Genomics track 2004 data set.

3.1.9 Limitations of the design

This study was designed to assess the impact of MeSH terms on search effectiveness in an interactive search environment. One limitation of the design was that participants were a self-selected group of searchers that may not be representative of the population. The interaction effects of selection biases and the experimental variable, i.e., the displayed MeSH terms, were another possible factor that limits the generalizability of this study (Campbell & Stanley, 1963). The use of relatively technical and difficult

search topics in the interactive search environment posed threat to external validity, since those topics might not represent typical topics received by medical librarians in practice.

3.2 Participant characteristics

As revealed in our review of related work, the user characteristics of domain knowledge and search training have been prominent in user behavior research in part because these characteristics are believed to be critical for conducting successful searches. We assumed that domain knowledge is primarily based on level of education, whereas search training comes from formal courses in relation to online searching. We measured the level of these two variables by the number of formal courses taken (see Appendix E) and distinguished four kinds of searchers: (1) Search Novices (SN); (2) Domain Experts (DE); (3) Search Experts (SE) and (4) Medical Librarians (ML).

To ensure that the participant demonstrates the level of domain knowledge as expected, the participant was instructed to rate the relevance of two documents for each assigned search topic. These two documents were randomly selected from the pool of relevance judged documents; one was ‘definitely relevant’ and the other was ‘not relevant’. The order of presentation was also randomized. The search topic judgment, subsequently termed comprehension test, was intended to ascertain that DEs and MLs demonstrate sufficient knowledge to understand technical search topics.

The participant profile overall satisfied the requirement of four kinds of searchers as specified by design. However, MLs did not reach a high level of biomedical knowledge as we expected. This verification of the participant’s level of domain knowledge has increased the internal validity of this study for the investigation of the impact of user characteristics.

3.2.1 Domain knowledge

DEs generally had the most biomedical knowledge as suggested by the large number of undergraduate (median = 15) and graduate (median = 7.5) levels of courses taken, followed by MLs. However, MLs' domain knowledge was much lower than that of the DEs and their biomedical knowledge primarily came from undergraduate courses (Figure 3-10 and 3-11). The DE searchers came from the subfields of Cancer Biology, Biochemistry & Molecular Biology, Chemical Biology, Neuroscience, Pharmacology, and Computational Biology & Molecular Biophysics.

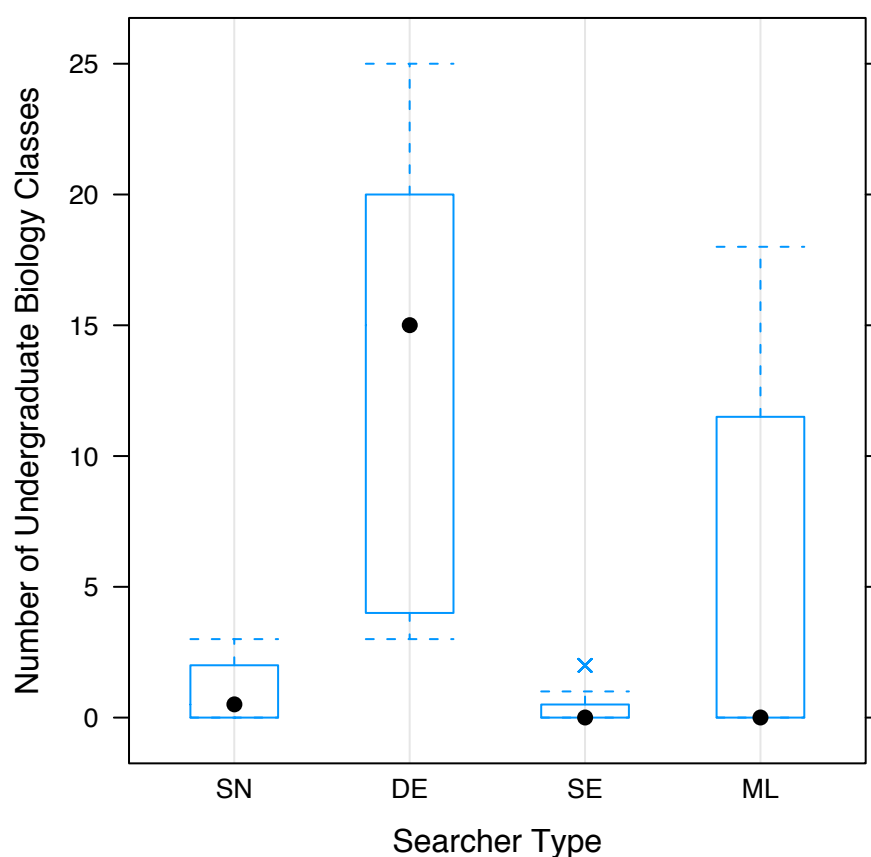


Figure 3-10 Box plot of undergraduate level biology knowledge by searcher type

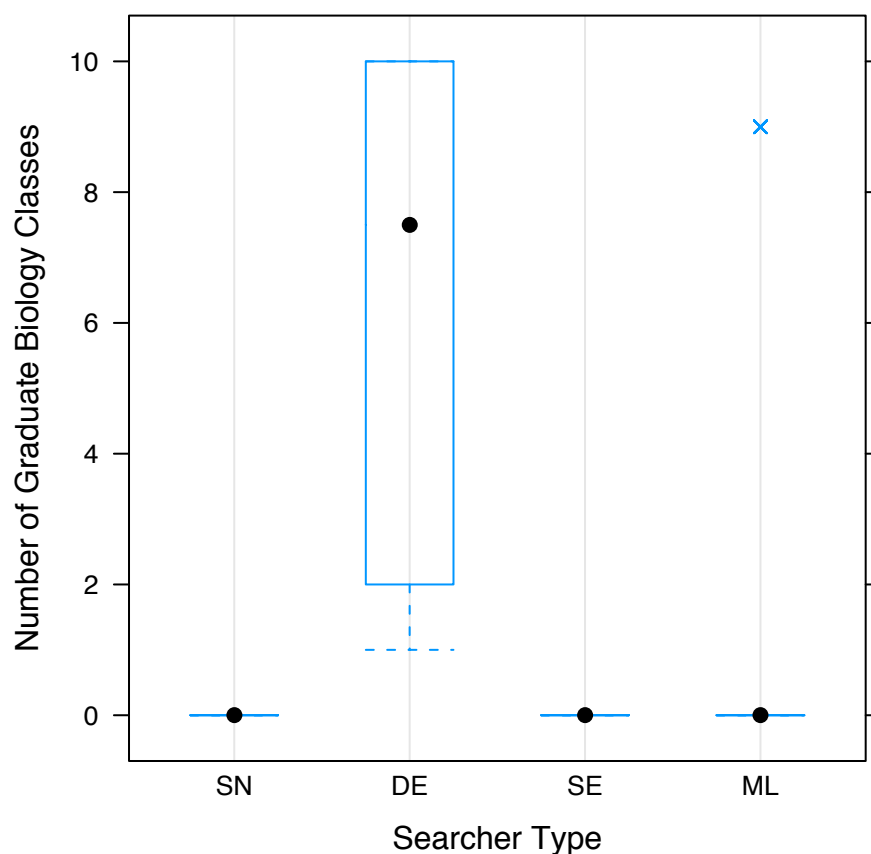


Figure 3-11 Box plot of graduate level biology knowledge by searcher type

Our results from comprehension test indicated that DEs demonstrate significantly better understanding of search topics than SEs do, whereas SNs and MLs fall between these two types of searchers. The correctness of judgment was composed of the following three categories: (1) both correct; (2) one correct and (3) none correct or not sure. There is a statistically significant relationship between searcher types and categories of correctness of judgment (Fisher's Exact Test, $p = .008$, $p < .05$). Note that the relevance judgment from the dataset is fairly reliable with Kappa statistic value of 0.51 for inter-judge agreement (Hersh, Bhupatiraju, Ross, Roberts, Cohen, & Kraemer, 2006; cf.

Saracevic, 2006). These results allowed us to conclude that DEs have significantly higher level of biomedical knowledge than SEs in this study. Overall, participants did not perform well in comprehension test and the results confirmed that levels of education are good indicators of domain knowledge.

But for each assigned search topic all types of searchers were only able to judge correctly both documents at 40-50% (Table 3-2). It is speculated that because of the diverse and deep coverage of assigned topics (see Appendix F for selected search topics), some genomic topics might be out of participant's area of expertise, or need additional time for research. In post-search interviews, some MLs also commented that these genomic topics were especially challenging because of the rapid development in this field and it was difficult to identify different names for a specific gene.

Table 3-2 Correctness of comprehension test by searcher type

Searcher Type	Correctness of Comprehension Test			
	Both Correct	One Correct	None Correct or Not Sure	Total
SN	37 (57.8%)	19 (29.7%)	8 (12.5%)	64 (100.0%)
DE	35 (54.7%)	27 (42.2%)	2 (3.1%)	64 (100.0%)
SE	28 (43.8%)	20 (31.3%)	16 (25.0%)	64 (100.0%)
ML	26 (40.6%)	28 (43.8%)	10 (15.6%)	64 (100.0%)
Total	126 (49.2%)	94 (36.7%)	36 (14.1%)	256 (100.0%)

Note. SN = Search novices; DE = Domain experts; SE = Search Experts; ML = Medical Librarians

3.2.2 Search training

Search training, measured by formal training in online searching course, suggested that MLs have participated in the largest number of online searching classes (median = 8.5), followed by SEs (median = 1) (Figure 3-12). Most DEs and SNs had no formal search training. MLs also had the most experience using MeSH terms. None of the SNs and DEs had used MeSH terms before they participated in the study (Table 3-3). As

would be expected MLs also had the most professional experience among the four types of searchers.

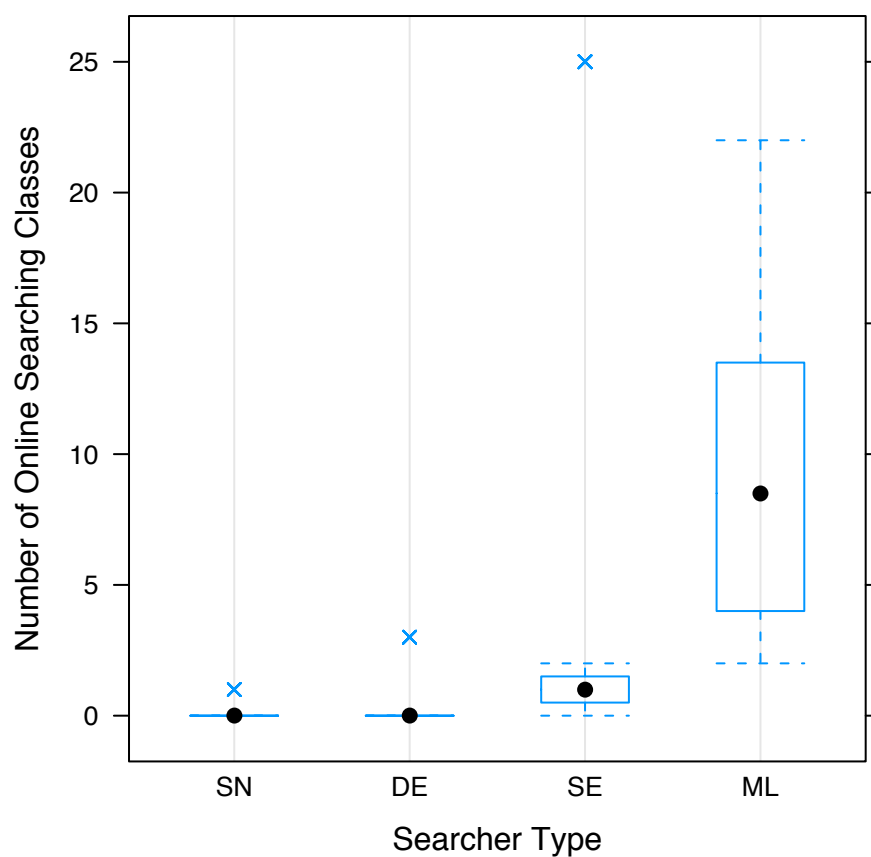


Figure 3-12 Box plot of search training by searcher type

Table 3-3 Amount of MeSH use experience

Searcher Type	Amount of MeSH Use Experience				Total
	None	A little	Some	A lot	
SN	8	0	0	0	8
DE	8	0	0	0	8
SE	4	3	1	0	8
ML	0	0	2	6	8
Total	20	3	3	6	32

Note. SN = Search novices; DE = Domain experts; SE = Search Experts; ML = Medical Librarians

3.2.3 Demographic variables

The participant's demographic profile revealed essential differences among the four types of searchers in terms of gender, native language and age variables. More specifically, MLs were mostly female native English speakers aged above 35, whereas nearly all DEs were male non-native English speakers aged between 25 and 35 (Table 3-5). Overall, we successfully recruited participants distinguished by different levels of domain knowledge and search training.

Table 3-4 Gender, native language and age by searcher type

Searcher Type	Gender		Native Language		Age			
	M	F	English	Other	18 – 25	25 – 35	35 – 45	> 45
SN	4	4	4	4	5	1	1	1
DE	7	1	0	8	0	8	0	0
SE	3	5	5	3	0	2	3	3
ML	1	7	7	1	0	1	2	5

Note. SN = Search Novice; DE = Domain Expert; SE = Search Expert; ML = Medical Librarian

In conclusion, we have successfully recruited different kinds of participants distinguished by level of domain knowledge and search training. As revealed by academic training in the biomedical domain and comprehension test, DEs demonstrated significantly higher level of domain knowledge than SEs; SNs and MLs' level of domain knowledge was between DEs and SEs.

MLs did not have much academic training in biology as we expected. Although we would have liked to say that MLs are SEs *and* DEs, their scientific knowledge is much less than the DEs. Furthermore, the fact that most MLs work in the medical field rather than biology may be a factor in their comprehension of genomic topics.

3.3 Data Analysis

To assess the usefulness of displayed index terms in interactive information seeking, we measured the search performance in terms of standard precision and recall measures, the characteristics of query terms and the impact of user perceptions about displayed terms usefulness and search task difficulty on search performance. All the statistical analysis was conducted using R software (“R Development Core Team”, 2008) and associated packages (Morales, 2009; Sarkar, 2009; Warnes, 2008a, b) because it provides a wide range of statistical techniques and well-designed graphics.

3.3.1 User search performance

We measured search performance by precision and recall measures, using the relevance judgments made for TREC. To calculate the user search performance, we used Perl scripts to process the search logs recorded on the experimental system when the participant finished each assigned search topic.

To assess participant’s performance for the monetary incentive, we used the trec_eval program and the relevance judgment from TREC Genomic track dataset to measure effectiveness (Buckley, 1999; “TREC 2004 Genomics Track document set,” 2005), when the searcher was done with each search topic.

Given that this study is based on a factorial experimental design and we are concerned with the effects of system, searcher and topic, we constructed a linear fixed-effects model to fit the data. The Graeco-Latin square design controlled three sources of variation: four types of searchers, two versions of a system and 10 search topic pairs. We considered the following model for

$$y(i,j,k) = m + s(i) + r(j) + t(k) + e(i,j,k)$$

where

$y(i,j,k)$ = precision/recall for system i , searcher j , search topic pair k

m = the mean precision/recall for the search results

$s(i)$ = effect of system i , where $i = 1$ (*MeSH-*), 2 (*MeSH+*)

$r(j)$ = effect of searcher j , where $j = 1$ (*SN*), 2 (*DE*), 3 (*SE*), 4 (*ML*)

$t(k)$ = effect of search topic pair, where $k = 1$ to 10 search topic pairs

$e(i,j,k)$ = the random error for observation $y(i,j,k)$

We performed square root data transformations on the precision and recall scores to approximate normal distributions (Fox, 1997; Hull, 1993; Tague-Sutcliffe, 1992, p. 485) for satisfying the requirement of analysis of variance. In addition to visual inspections of quantile comparison plots, the Shapiro-Wilk test of normality allowed us to conclude that we cannot reject the null hypothesis that the transformed data is from a normal distribution.

The approach of factorial design and analysis allowed us to separate the effects of systems, searchers and search topics, with a relatively small number of participants. The transformation on the precision and recall scores ensured the requirement of analysis of variance within a linear fixed-effects model.

3.3.2 Query terms characteristics

To explore the search processes, we calculated and compared the query terms by systems and searcher types, and considered query terms characteristics: (1) the number of terms per search session (token), (2) the number of unique terms per search session (type), (3) the number of queries per search session and (4) the number of terms per query (query length). The result of query term characteristics will give insight into

searcher's representation of biologist's information needs and the accuracy of search results.

The analysis of query terms involved the basic unit of analysis and identification of unique search terms that reflect searcher's knowledge about the search topic. The basic unit of analysis was any search term issued by searchers. More specifically, any word separated by blank space, including prepositions and articles, was counted as a search term. To ensure the consistency in identifying unique search terms, we devised and followed several rules (see Figure 3-13 for sample query logs):

1. The unit of analysis is the term or phrase within the brackets in query logs. The search fields and Boolean operators are not considered;
2. The terms or phrases are not considered unique if there are only singular/plural differences;
3. The terms or phrases are not considered unique, if the difference comes from the case sensitivity and stemming search options specified in the search system;
4. The terms or phrases are not considered unique, if the words have different meanings, such as *ethnically*, *ethnical* and *ethnic*.

```
<S03, t42>
[genes ]:KE
[genes ]:TI
[genes#si ]:AB AND [chromosomes#s ]:AB AND [translocation#s ]:AB
[gene#si ]:TI AND [chromosome#si ]:TI AND [translocation#si ]:TI

<S15, t12>
[signal transducin ]:TX AND [smad4 targets ]:TX AND [gene expression ]:TX AND
[skin ]:TX
[signal transducin ]:TX AND [Smad4 targets ]:TX AND [gene expression ]:TX AND
[skin ]:TX
[signal transducin ]:TX AND [Smad4 ]:TX AND [gene expression ]:TX AND [skin ]:TX
```

```

[signal transducing ]:TX AND [Smad4 ]:TX AND [gene expression ]:TX AND [skin ]:TX
[signal transduction ]:TX AND [Smad4 ]:TX AND [gene expression ]:TX AND [skin ]:TX
[signal transduction ]:TX AND [signaling network ]:TX AND [gene ]:TX AND [skin ]:TX
[signal transduction ]:TX AND [signal network ]:TX AND [gene ]:TX AND [skin ]:TX
[signal transduction pathway ]:TX AND [mouse ]:TX AND [knock ]:TX AND [skin ]:TX
[signal ]:TX AND [mouse ]:TX AND [knock ]:TX AND [skin ]:TX
[signal ]:TX AND [mice ]:TX AND [knock ]:TX AND [skin ]:TX
[Smad4 ]:TX AND [mice ]:TX AND [knock ]:TX AND [skin ]:TX

```

Note. <S03, t42> indicates query logs from subject 03 searching topic 42; Search fields KE = MeSH terms, TI = Title, AB = Abstract, TX = Full records; For case sensitivity and stemming, #i means the search term is case insensitive and #s means the search term is stemmed.

Figure 3-13 Query log examples

An analysis of variance (ANOVA) was performed between the query terms characteristics and the system versions, the searcher types and the system versions and searcher type pairs. Tukey's HSD multiple comparisons were also conducted when searcher types make a difference in query terms characteristics. The results will reveal whether query term characteristics vary by system versions, searcher types and system-searcher interactions and provide additional information for interpreting search effectiveness.

3.3.3 User perceptions and search performance

To determine the relation between user perceptions and search performance, we performed logarithmic cross ratio analysis between the two variables. In the context of information retrieval experiment, this technique is particularly useful because it takes into account the self-selection of participants in the study and the skewed distribution of relevance score (Fleiss, Levin, & Paik, 2003; Saracevic, Kantor, Chamis, & Trivison,

1988). For user perceptions we considered the usefulness of displayed query terms (MeSH+ and MeSH– versions) and search task difficulty. User search performance was considered in terms of the precision and recall score (search effectiveness) and the time spent (search efficiency). The same technique was also used to assess the relation between user characteristics and search effectiveness.

3.4 Summary

This controlled information retrieval experiment was designed to assess the usefulness of MeSH terms for users with different levels of biomedical domain knowledge and search training. We observed four different kinds of information seekers using an experimental information retrieval system: (1) search novices (SN); (2) domain experts (DE); (3) search experts (SE) and (4) medical librarians (ML). The information needs were a subset of the topics originally created for Text REtrieval Conference (TREC) Genomics Track 2004. All of these topics were relatively difficult; we used domain experts to help identify the most comprehensible to a general audience. Effectiveness of retrieval was based on the relevance judgments provided by TREC. Participants searched either using a version of the system in which MeSH terms were displayed (MeSH+) or another version in which they had to formulate their own terms (MeSH–).

The internal validity of this design was enhanced by specifically considering several aspects: We devised an incentive system to consider the possible sampling bias of searchers' motivational characteristics in experimental settings. Besides levels of education, participants' domain knowledge was evaluated by a topic understanding test. The variability of search topics was alleviated by using a relatively large number of

search topics by experimental design. Selected search topics were intelligible in consultation with domain expert and medical librarian. A concept analysis form was used to help searchers recognize potentially useful terms. The reliability of relevance judgment sets was ensured by additional analysis of top 10 search results from our human searchers.

In the next chapter, we will report our findings of user search performance and search efficiency under these controlled conditions.

CHAPTER 4 RESULTS

This chapter reports the results of search performance measured by search effectiveness in terms of precision and recall measures, and search efforts in terms of time spent for the use of different versions of a system and for different kinds of users. We then examine query reformulations in terms of query term characteristics and search effectiveness. Finally, we consider the relationship between user perceptions of search tasks and search performance.

The genomics search topics used in the study were relatively technical in nature and required additional time for research, as suggested by searchers' overall comments on search tasks. In this chapter, we report search results; they shall all be interpreted in light of this perspective in chapter five.

4.1 Overall Use of MeSH Terms

This study was designed to assess the impact of MeSH terms on search effectiveness by different types of searchers. The search logs revealed that participants overall did use MeSH terms during search processes when they used MeSH+ version (Table 4-1). Searchers specified MeSH terms as search field in 39.1% of all searches.

Table 4-1 Use of MeSH terms search field in MeSH+ version

	MeSH Terms Search Field Use in Searches		
Searcher Type	Yes	No	Total
SN	2 (6.3%)	30 (93.8%)	32 (100.0%)
DE	10 (31.3%)	22 (68.8%)	32 (100.0%)
SE	16 (50.0%)	16 (50.0%)	32 (100.0%)
ML	22 (68.8%)	10 (31.3%)	32 (100.0%)
Total	50 (39.1%)	78 (60.9%)	128 (100.0%)

Note. SN = Search novices; DE = Domain experts; SE = Search Experts; ML = Medical Librarians. For each searcher type, there are 32 searches in total (8 searchers × 4 topics =

32 searches.)

Further analysis suggested that there was a statistically significant relationship between searcher types and use of MeSH terms (Fisher's Exact Test, $p = 7.717\text{e-}07$, $p < .001$). Searchers' levels of search training were reflected in the use of MeSH terms; the more search training one had, the more likely one would use MeSH terms. After a brief training session, DEs were able to search with MeSH terms, even though they had not used MeSH terms before. So DEs easily learned to use MeSH terms and this is part of their search training. These results validated our experimental instruments and procedures.

4.2 Search Efficiency

The participants were very engaged with assigned search tasks. A density histogram of time spent by all searches with a superimposed theoretical normal curve showed an extremely high-density value of time within a range of 550-600 seconds (Figure 4-1). There was no significant difference in the time spent using MeSH+ or MeSH- versions (ANOVA, $F(1, 254) = 2.77$, $p = .10$, $p > .05$). However, the time spent by searcher types was statistically significant (ANOVA, $F(3, 252) = 3.47$, $p < .05$). Further analysis showed that DEs spent significantly more time than SEs (Tukey HSD, $p < .05$).

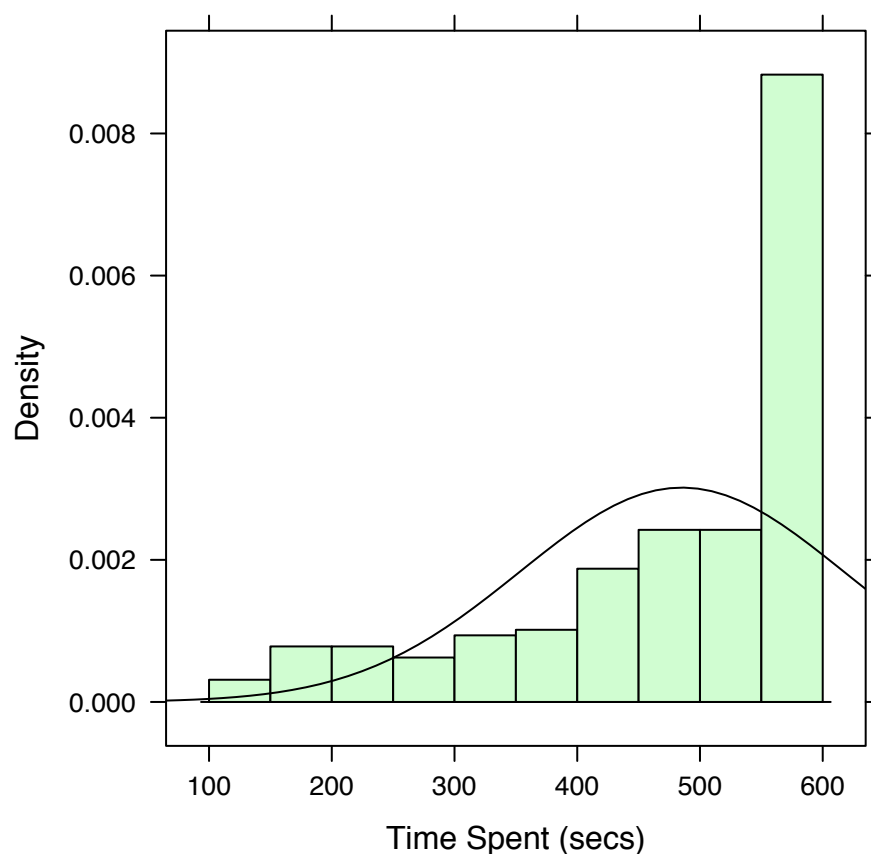


Figure 4-1 Histogram of time spent with normal density overlaid by all searches ($N = 256$)

We also found that the time spent is significantly different based on searcher type and system versions (Figure 4-2). More specifically, there were significant differences in the time spent by searcher types across system versions (ANOVA, $F(7, 248) = 2.41, p < .05$). Further analysis indicated that the time spent by DEs using MeSH+ was longer than SEs using MeSH- (Tukey's HSD, $p < .05$). We speculate that this is a result of the searcher groups' different level of expertise.

The greater amount of time may reflect at least these two factors: 1) Trained searchers found the topics difficult; 2) Trained searchers know how to persist in

searching, even when they were having a hard time. It is speculated that because of the relatively technical nature of search topics, DEs are able to be engaged in searching by drawing upon their domain knowledge no matter what kinds of search tools are offered. For trained searchers, they are still persistent in searching MeSH+ system when they are given difficult topics.

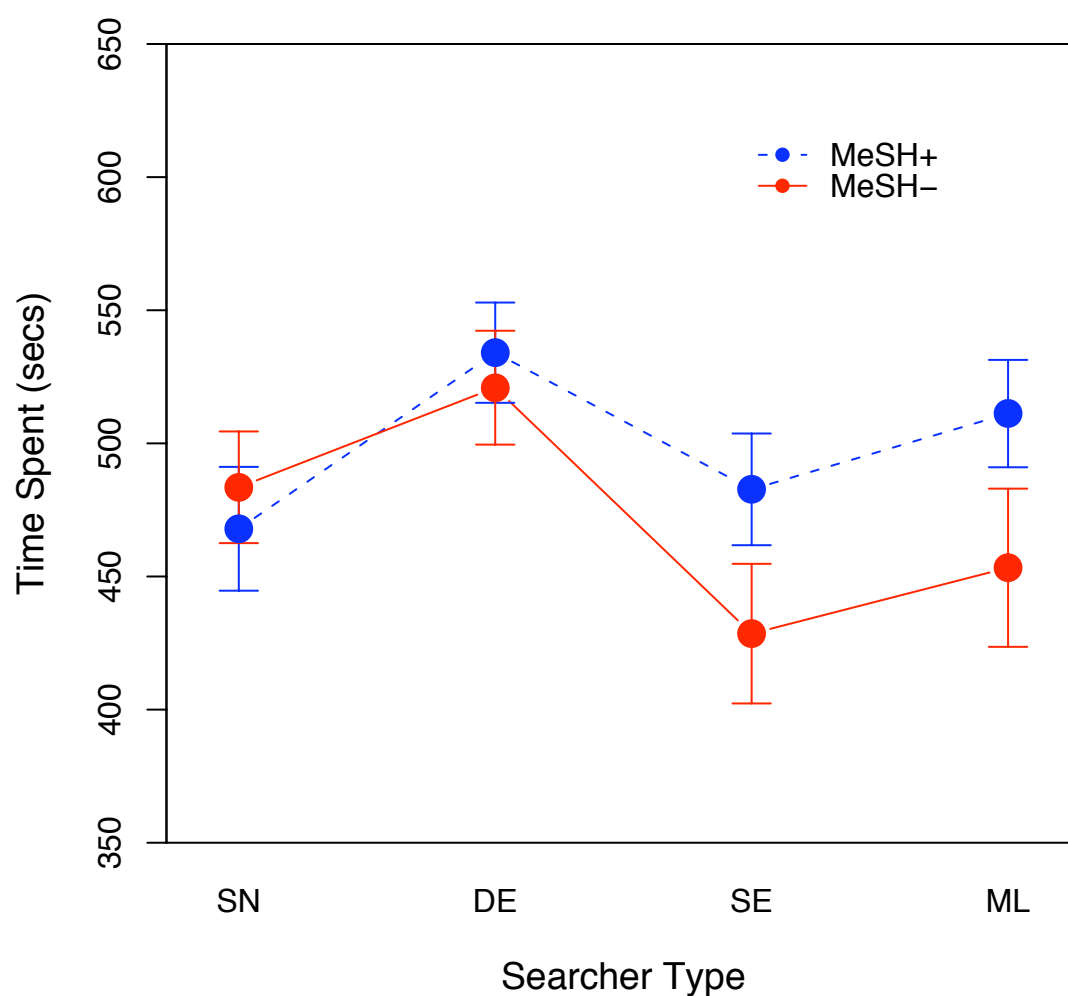


Figure 4-2 Line plot of the mean and standard error of time spent by searcher type and system version

4.3 Search Outcome

We measured search outcome in terms of precision and recall measures for search effectiveness and time spent for search efficiency. The overall result comparing MeSH+ and MeSH– versions suggested that there was no statistically significant difference between the two versions of the experimental system, in terms of both precision (ANOVA, $F(1, 254) = 0.01, p = .94, p > .05$) and recall (ANOVA, $F(1, 254) = 0.30, p = .58, p > .05$) measures. The hypothesis that queries using MeSH will get better results than queries not using MeSH thus is not supported. In chapter five, we will discuss the possible reasons for this result.

Different types of searchers obtained comparable results when we compared all search results, regardless of system versions. Search effectiveness by different types of searchers did not make statistically difference in terms of precision (ANOVA, $F(3, 252) = 1.86, p = .14, p > .05$) and recall (ANOVA, $F(3, 252) = 1.66, p = .18, p > .05$) measures. All four types of searchers were only able to achieve mean precisions of approximately between .30 and .40, and mean recalls between .15 and .23 (Table 4-2). This result showed that search tasks were difficult for all searchers.

Table 4-2 Search effectiveness by searcher types in terms of precision and recall measures

Searcher Type	Mean Precision	N	Mean Recall	N
SN	0.29	64	0.21	64
DE	0.40	64	0.15	64
SE	0.30	64	0.15	64
ML	0.35	64	0.23	64
Total	0.34	256	0.18	256

Note. SN = search novices; DE = domain experts; SE = search experts; ML = medical librarians

But when we compared search effectiveness of different types of searchers across system versions, we found a very strong effect of system version and searcher type in terms of precision measure (ANOVA, $F(7, 248) = 3.48, p = .001, p < .01$) (Table 4-3). In particular, there were highly significant differences in precision between DEs and SEs when they used the MeSH+ version (Tukey's HSD, $p < .01$) and between DEs' use of MeSH+ and SNs' use of MeSH- versions (Tukey's HSD, $p < .01$) (Figure 4-3). A statistical power analysis where significance and power levels were set at 0.01 and 0.80 produced a medium effect size ($ES = 0.28$) (Champely, 2007; Cohen, 1988).

Table 4-3 Search effectiveness by system version and searcher type in terms of precision and recall

	MeSH+			MeSH-		
Searcher Type	Mean Precision	Mean Recall	N	Mean Precision	Mean Recall	N
SN	0.36	0.21	32	0.23	0.20	32
DE	0.51	0.15	32	0.29	0.15	32
SE	0.21	0.16	32	0.38	0.13	32
ML	0.28	0.22	32	0.42	0.24	32
Total	0.34	0.19	128	0.33	0.18	128

Note. SN = search novices; DE = domain experts; SE = search experts; ML = medical librarians

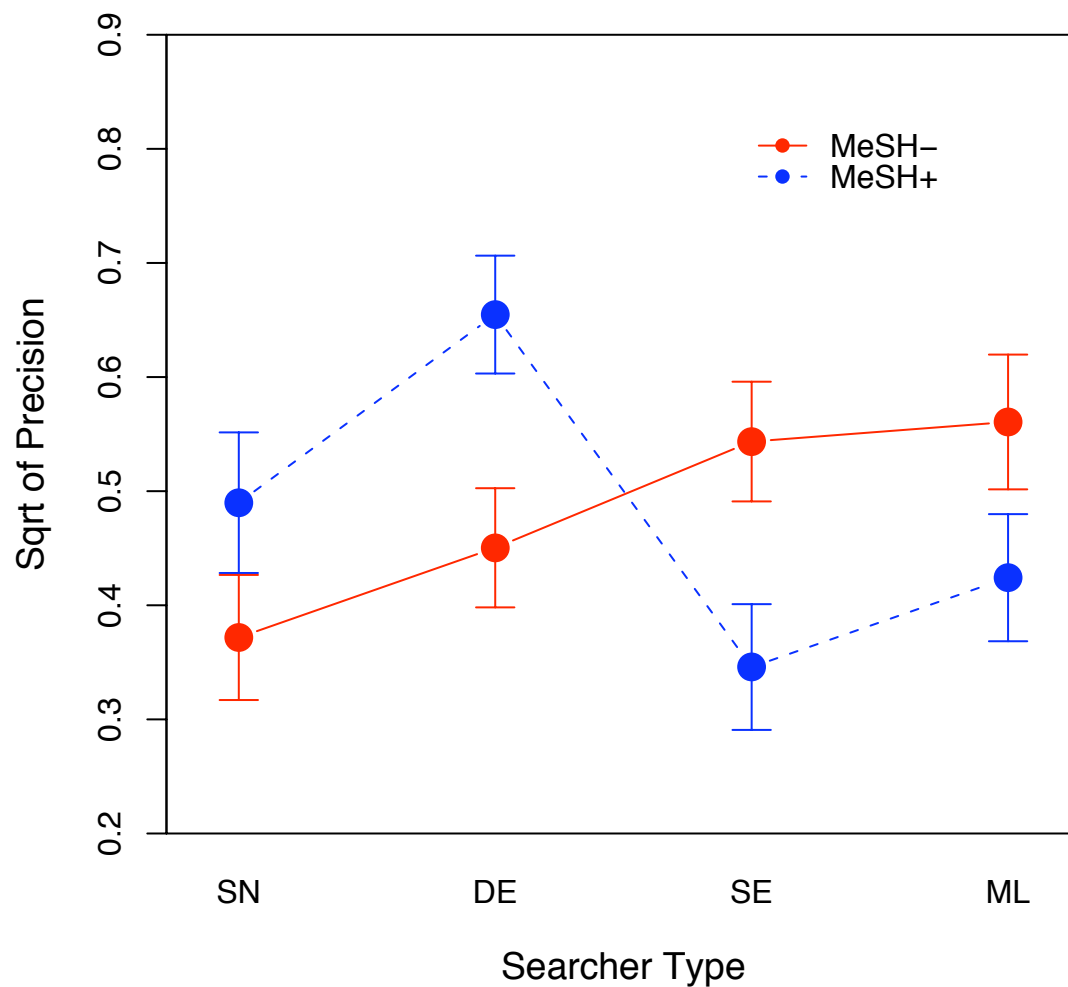


Figure 4-3 Line plot of the mean and standard error of square root of precision by searcher type and system version

In general, searchers with the most domain knowledge (DEs) were capable of obtaining significantly better search results than SEs with the help of MeSH terms. Trained searchers (SEs and MLs) were able to achieve better search results than untrained searchers (SNs and DEs) when MeSH terms were not offered, although the

increase in effectiveness was not statistically significant. Domain knowledge makes a big difference using MeSH terms in terms of precision.

These results represent a form of interaction in factorial design in which neither searcher type nor system version has any main effect, but in which the interactions are strong and definite (Figure 4-4). To distinguish between different kinds of interactions, we focus next on the features that compose different types of searchers: domain knowledge and search training. Because of the significant interaction effects in factorial design, we should be reserved about interpreting effects for generalizing purposes.

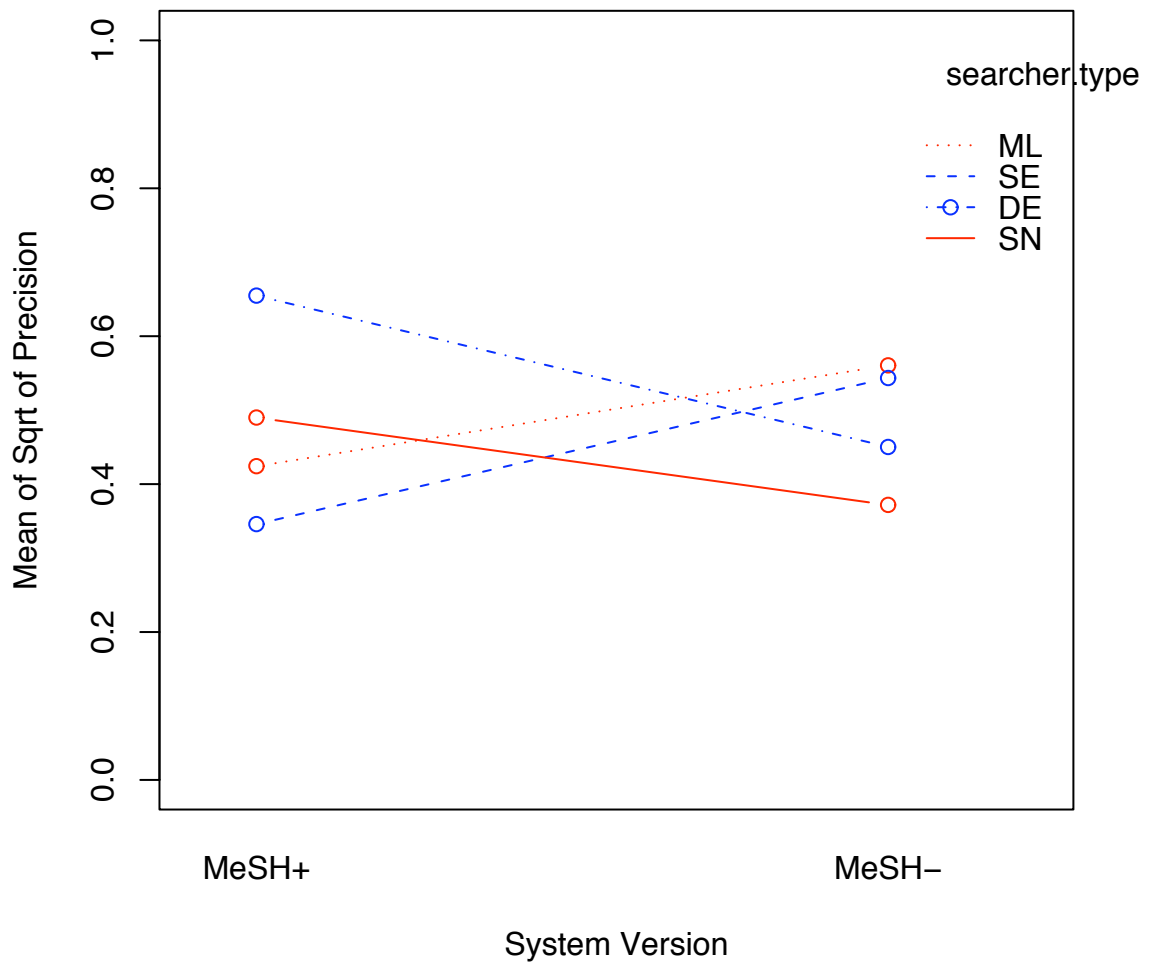


Figure 4-4 Interaction plot between searcher type and system version

The results reveal a significant main effect of domain knowledge ($F(1, 124) = 4.16, p < .05$) and highly significant interaction effects between domain knowledge and system version ($F(1, 124) = 14.45, p < .001$) in terms of the precision measure. Only the contrast between DEs and SEs in their levels of domain knowledge was considered because of the comprehension test result that MLs' level of domain knowledge was not as high as expected. Compared to SEs who had search training but presumably no advanced

biomedical knowledge, MLs had less academic training in the biomedical domain than expected. These results suggest that domain knowledge is crucial for searching technical topics, and that domain experts using MeSH terms can significantly enhance the precision of searches. On the whole, domain experts benefit the most from the use of MeSH terms.

Search training alone does not make a difference in terms of precision ($F(1, 252) = 0.35, p > .05$), but there are strong interaction effects between search training and system version ($F(1, 252) = 17.37, p < .001$). The level of search training across all searchers was considered for analysis ($N = 256$) because formal search training was widely different between trained and untrained searchers. One possible explanation is that searchers with low level of search training can search reasonably well, partly because the experimental system is equipped with state-of-the-art retrieval techniques and the search results are ranked by order of relevance. This is evidence that MeSH terms are not useful for people who are not domain experts. Later we will discuss why this might be and why we should not generalize from the result.

4.4 Comparison of Human and Computer Performance

We compared the human and computer performance by search topic because of the large variability in topic performance, as shown in this study (ANOVA, $F(9, 242) = 5.19, p < .001$). The search effectiveness obtained by computer systems came from the original TREC Genomics Track 2004 dataset in which we had 47 different runs from 27 research groups (Hersh et al., 2006). Our experimental design resulted in 16 or 20 searches per topic from 32 searchers.

A comparison of the average search performance, using MAP, P10 and P100 measures indicated that machine runs are consistently better than human searches. More specifically, there was significant difference between the human and the machine runs in terms of the MAP measure (two-tailed paired t-test, $t(19) = -6.29, p < .001$, mean of the differences = 12.5%) (see Figure 4-5), the P10 measure ($t(19) = -3.76, p < .01$, mean of the differences = 13.6%) (see Figure 4-6) and the P100 measure ($t(19) = -6.23, p < .001$, mean of the differences = 17.5%) (see Figure 4-7).

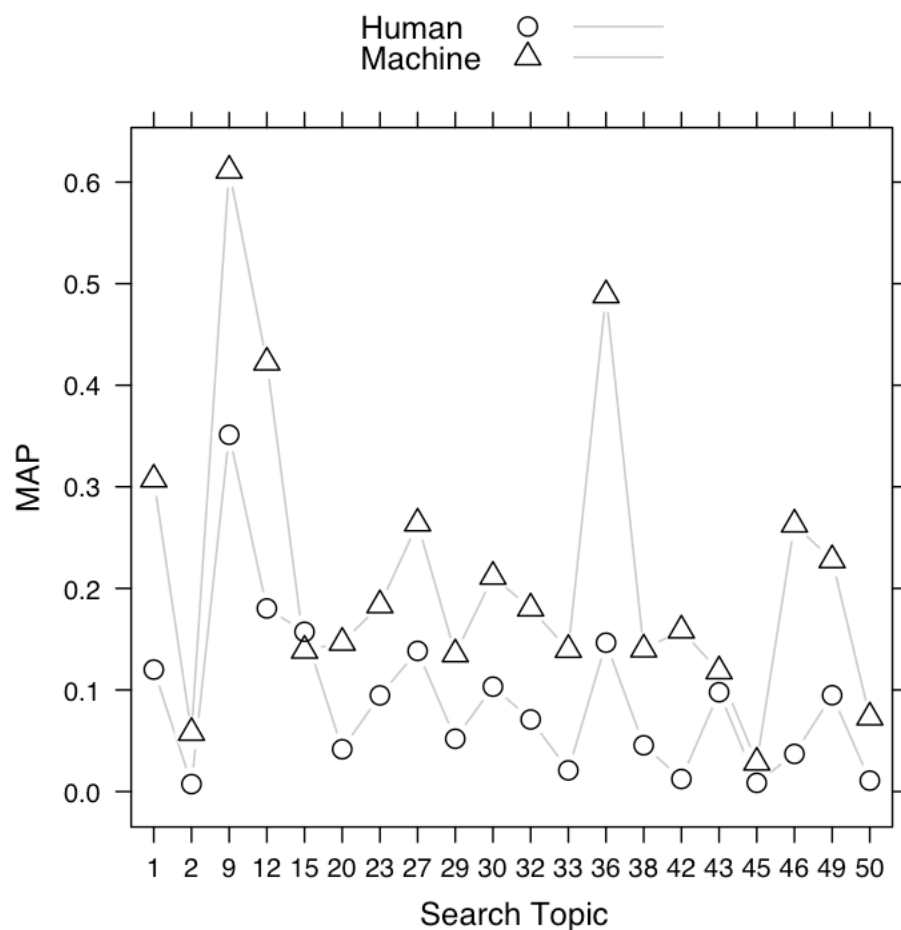


Figure 4-5 Plot of MAP (mean average precision) by search topic between human and computer searches

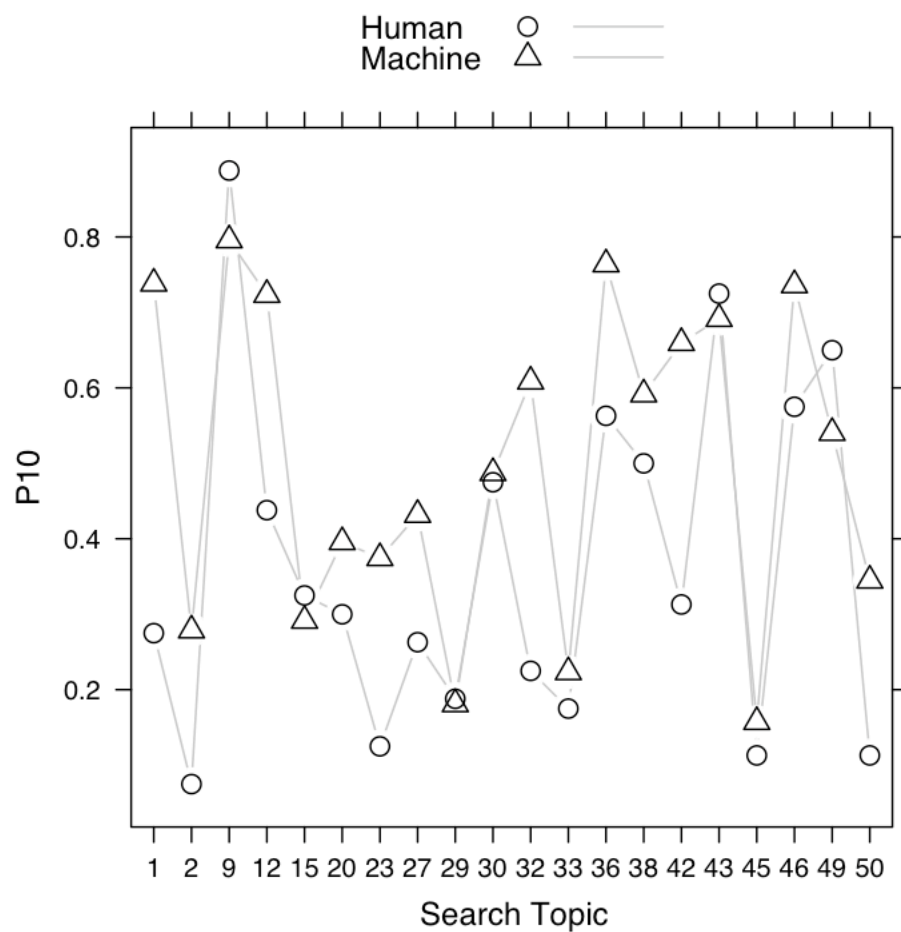


Figure 4-6 Plot of P10 (precision after 10 documents retrieved) by search topic between human and computer searches

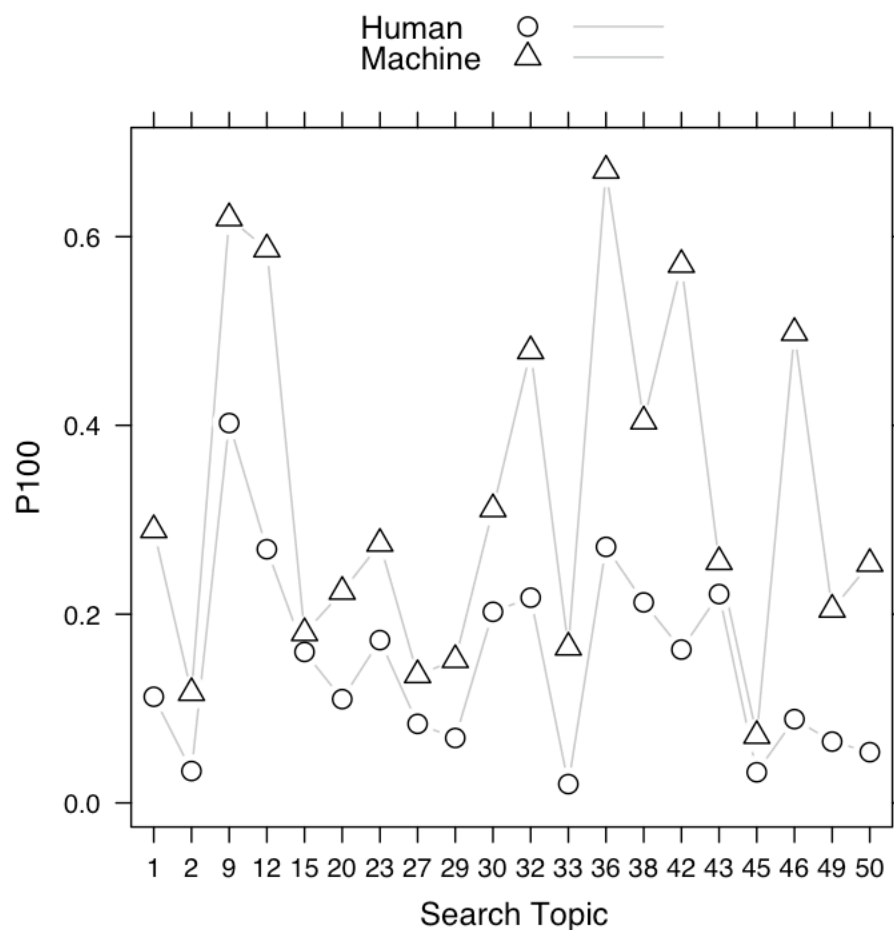


Figure 4-7 Plot of P100 (precision after 100 documents retrieved) by search topic between human and computer searches

One possible explanation for the considerably better performance archived by computer runs than human searches is that because these search topics were originally designed for comparing search effectiveness of different retrieval techniques, systems would do well in ad hoc search tasks. The systems performed reasonably well since they were specially trained for obtaining high precision results regardless of the intelligibility of topics. For human searchers these topics were still too difficult particularly for those who had no technical knowledge in the biomedical domain, although we considered this factor in our selection of search topics. Different variants of gene names and limited time

for searching made search tasks difficult, as human searchers reported. We will interpret this finding in light of the differences in IR batch-mode and user experiments in the next chapter.

4.5 Query Terms

To understand the search processes and gain more insight into the search outcome, we measured and compared the quantity and distinctiveness of query terms and the queries issued by different types of searchers. Our overall results indicate that searcher types made significant differences in all the measures, especially when MeSH+ version was offered. More specifically, DEs issued significantly more queries, and thus used more terms, than SNs did. And MLs used significantly more unique terms than SN did. We speculate that different search behaviors manifested by query terms may be affected by domain expertise and search training.

4.5.1 Number of terms per search session (tokens)

Searcher types made a significant difference in the total number of terms per search session (ANOVA, $F(3, 252) = 3.82, p < .05$). More specifically, DEs used significantly more terms than SNs did (Tukey's HSD, $p < .05$). There was also a significant interaction effect between searcher type and system version (ANOVA, $F(7, 248) = 2.42, p < .05$). DEs used significantly more terms than SNs did when they searched MeSH+ version (Tukey's HSD, $p < .05$) (Figure 4-8). This difference might be attributed to searchers' level of domain expertise in which DEs were capable of drawing upon their domain knowledge and further explore the search topic, while SNs did not have sufficient knowledge to make good use of MeSH terms.

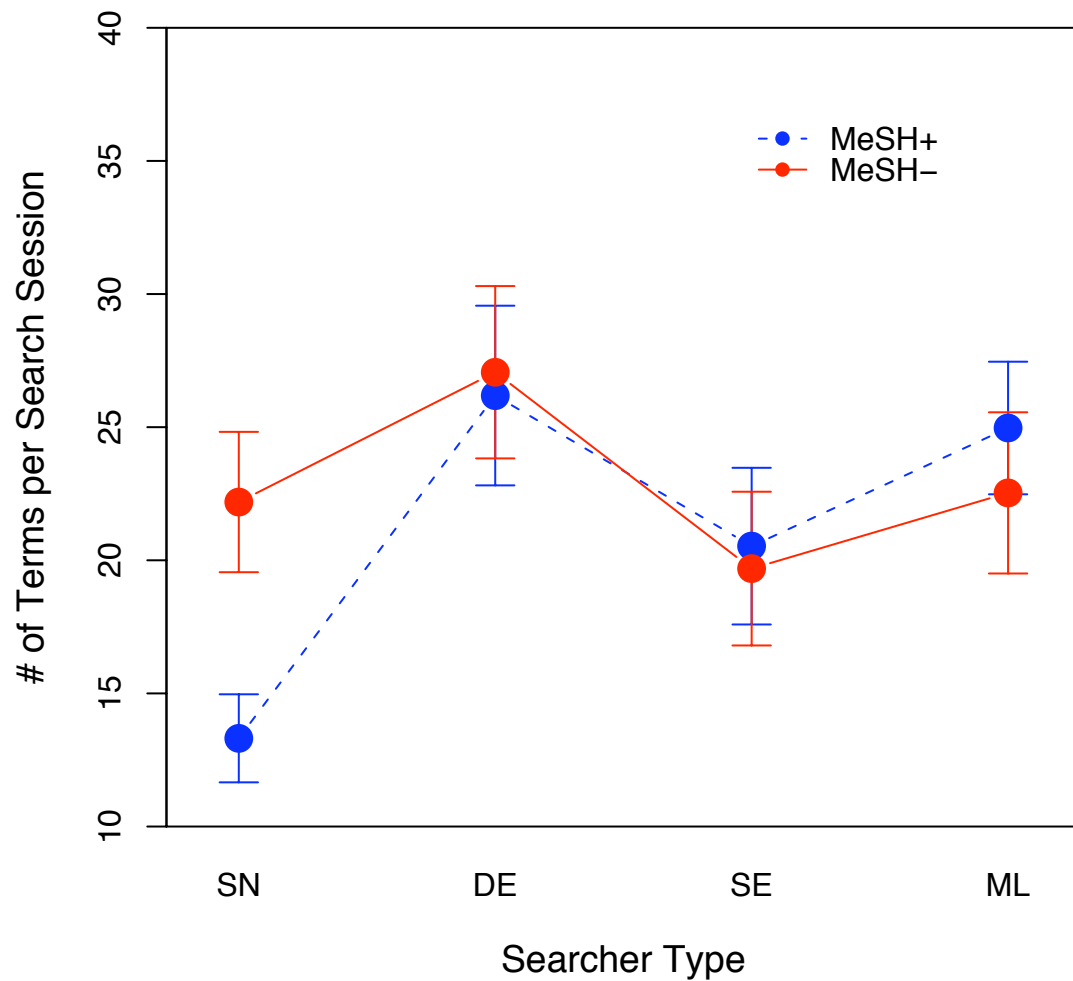


Figure 4-8 Line plot of the mean and standard error of terms per search session by searcher type and system version

4.5.2 Number of unique terms per search session (types)

Searcher types made a significant difference in the number of unique terms per search session (ANOVA, $F(3, 252) = 3.16, p < .05$). More specifically, MLs used significantly more unique terms than SNs did (Tukey's HSD, $p < .05$). There was also a significant interaction effect between searcher type and system version ($F(7, 248) = 2.06$,

$p < .05$) (Figure 4-9). MLs used significantly more unique terms than SNs did when they were offered MeSH+ version (Tukey's HSD, $p < .05$).

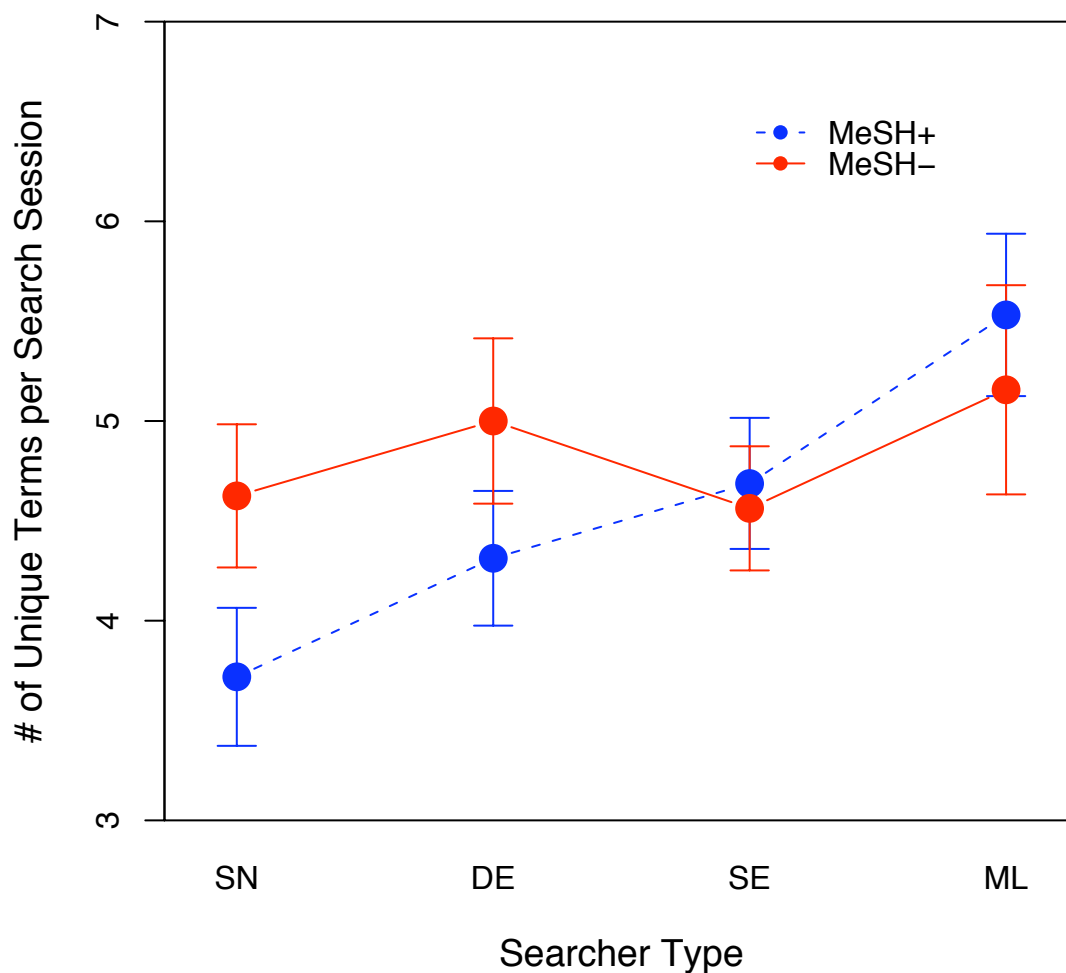


Figure 4-9 Line plot of the mean and standard error of unique terms per search session by searcher type and system version

The greater number of terms and unique terms used by DEs and MLs respectively may reflect at least two factors: 1) DEs were able to draw on their domain knowledge in searches; 2) MLs were capable of using MeSH terms to expand their scope of finding potentially useful terms. SNs, without substantial biomedical knowledge and search

training, used the least number of terms and unique terms when MeSH terms were offered. This is another indication that MeSH terms were especially not helpful for SNs who rarely specified the MeSH terms search field (see Table 4-1). SEs searched so similarly with and without MeSH terms. The data is hard to interpret because half SEs had previous experience using MeSH terms (see Table 3-3).

4.5.3 Number of queries per search session

The number of issued queries is an indication of search effort. Searcher type made significant differences in the number of issued queries (ANOVA, $F(3, 252) = 4.81, p < .01$). But system version alone did not make a difference (ANOVA, $F(1, 254) = 0.44, p > .05$). More specifically, DEs used significantly more queries than both SNs (Tukey's HSD, $p < .05$) and SEs (Tukey's HSD, $p < 0.05$) did. Further analysis indicated that DEs issued significantly more queries than SNs when they searched MeSH+ version (Tukey's HSD, $p < .05$) (Figure 4-10).

The significant difference between DEs and SNs in the number of issued queries might reflect the intrinsic difficulty of assigned search topics that required a lot of effort even for DEs, or it might be that DEs were more interested or that they were more motivated because they understood the results.

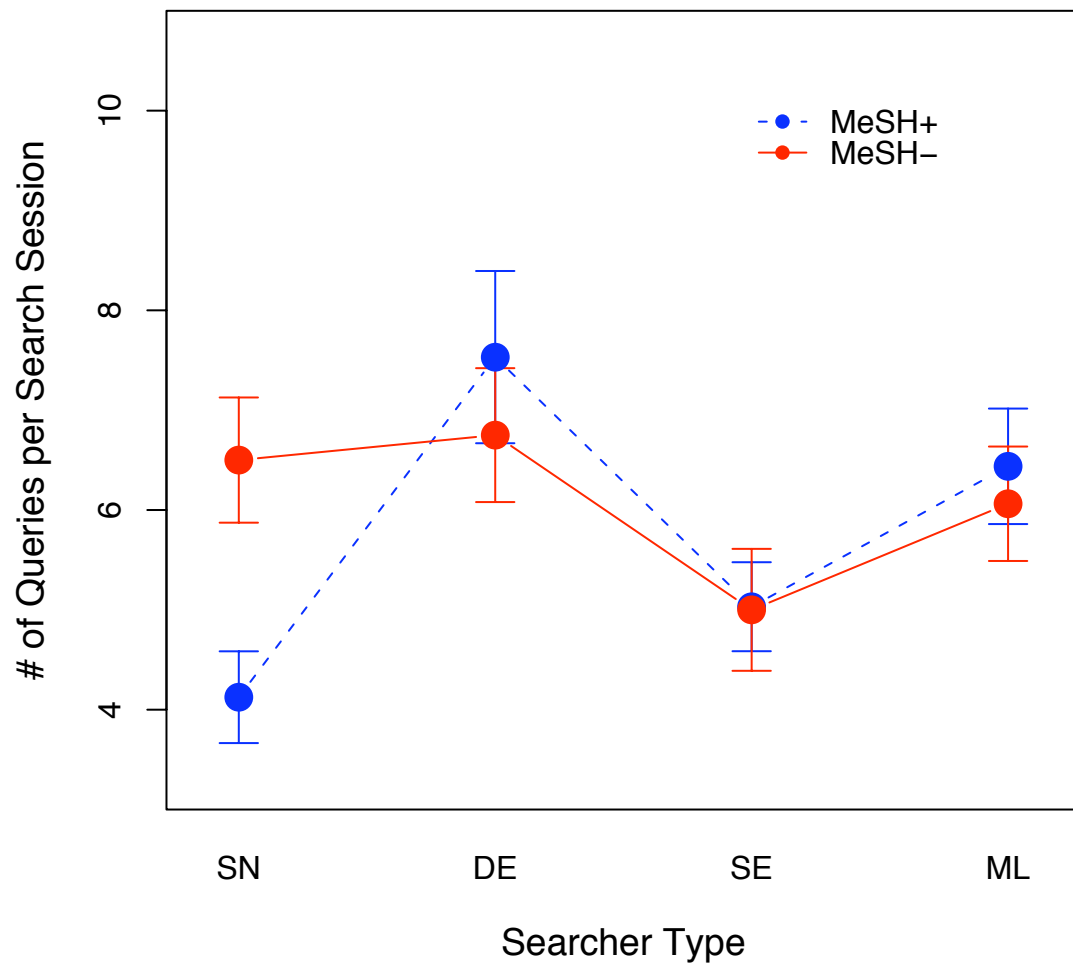


Figure 4-10 Line plot of the mean and standard error of queries per search session by searcher type and system version

4.5.4 Number of terms per query (query length)

Query length in best-match IR systems is positively correlated with search results in both batch-mode evaluation and interactive search environment (e.g., Belkin et al., 2003; Xu & Croft, 1996). Our controlled user experiment using a Boolean-based IR system with ranked search results provided another interactive search environment for testing.

Our result indicated that neither searcher type nor system version made a difference in query length. There was no significant difference in query length by searcher type (ANOVA, $F(3, 252) = 2.59, p > .05$) and system version (ANOVA, $F(1, 254) = .06, p > .05$) (Figure 4-11). MLs had the longest queries in using MeSH terms, but the differences were not statistically significant.

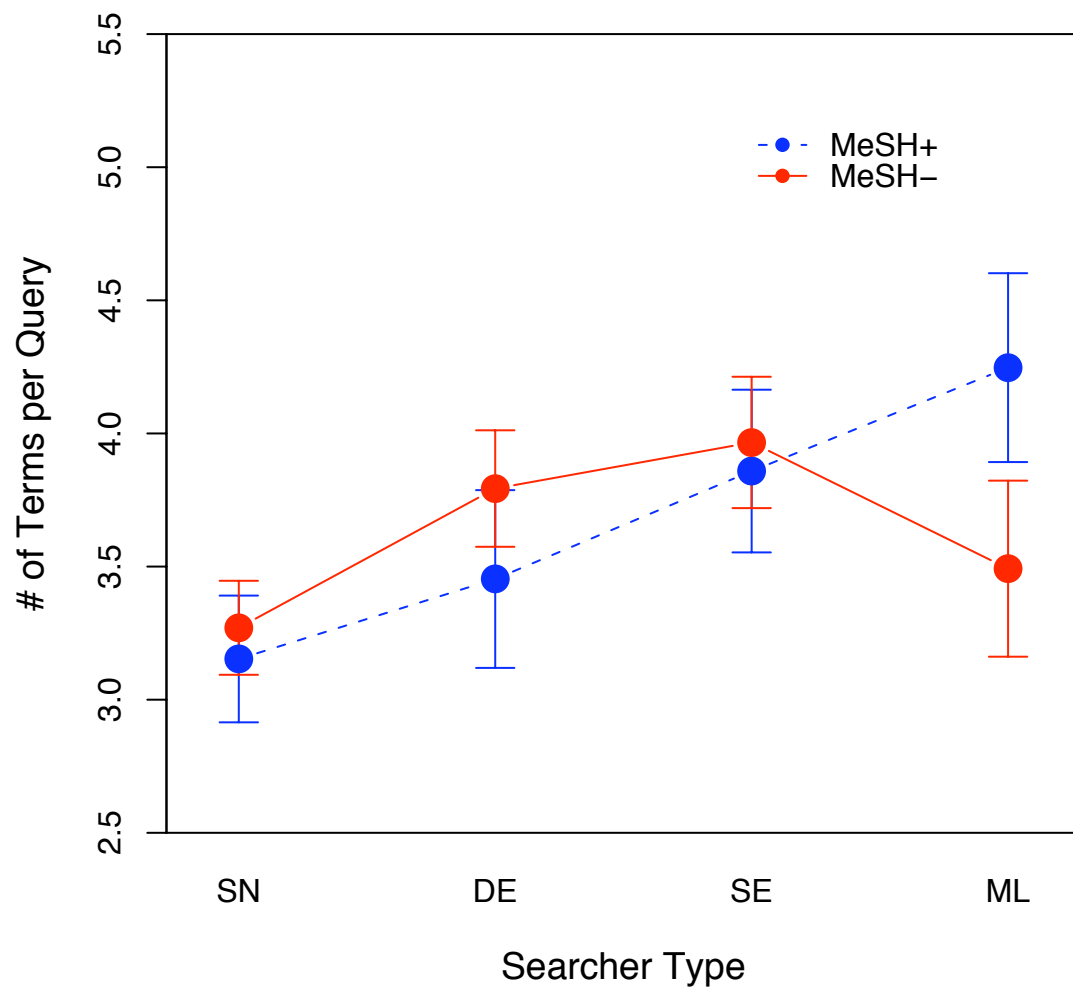


Figure 4-11 Line plot of the mean and standard error of terms per query by searcher type and system version

Overall, searcher type made significant differences in the tokens and types of query terms and issued queries (see Table 4-4 for a summary of results). These differences were particularly strong when MeSH+ version was offered, but system version alone did not make such differences. DEs used significantly more queries and query terms than SNs when MeSH+ version was offered. This is again evidence of usefulness of MeSH terms for DEs.

Table 4-4 Summary of query terms results

	<i>Tokens</i>	<i>Types</i>	<i>Queries</i>	<i>Query Length</i>
System Version	No sig. diff.	No sig. diff.	No sig. diff.	No sig. diff.
Searcher Type	DE >> SN	ML >> SN	DE >> SN	No sig. diff.
Searcher Type-System Version	DE (MeSH+) >> SN (MeSH+)	ML (MeSH+) >> SN (MeSH+)	DE (MeSH+) >> SN (MeSH+)	No sig. diff.

Note. >> means better at .05 level of significance; No sig. diff. = no significant difference; Sig. diff. = significant difference at .05 level; DE = domain experts; SN = search novices; ML = medical librarians; MeSH+ = MeSH+ version search system.

This might reflect that DEs were more interested and motivated because they understood the search results, or that the search tasks were difficult even for DEs. Also MeSH terms helped since DEs understood the terminologies. The fact that MLs used significantly more unique terms than SNs might be attributed to their search training and MeSH terms use experience. In these circumstances query length wasn't enough to improve MLs' search results.

4.6 Query Reformulations

Searches with query reformulations are indicative of user's intent to obtain more or better search results. The improvement of results as correlated with query reformulations was calculated by comparing results of the first and last search. The

overall result revealed that searches with query reformulations obtain better search results in terms of the precision, but not in terms of the recall measure.

There were statistically significant differences in terms of the precision measure (two-tailed paired *t*-test, $t(228) = -4.98, p < .001$). The difference in means was not equal to zero, with mean of differences equal to $-.131$, meaning that query reformulations improved the precision score by 13.1%. For the recall measure, there was no significant difference (two-tailed paired *t*-test, $t(228) = -1.54, p > .05$). Thus, the hypothesis that query reformulations will obtain better search results than initial searches is supported in terms of the precision measure. But query reformulation did not improve the recall score probably because searchers were motivated by our incentive system based on the precision of top 10 search results

Not all types of searchers benefited from query reformulations in the study. Trained searchers (SEs and MLs) and domain experts (DEs) were able to improve the precision score by 12.0-18.6% (Table 4-6). Given the relatively difficult search topics and the least number of queries issued by SNs, their results did not get better—they did not know how to improve queries. This may be because of limited searching skills and/or limited domain knowledge.

Table 4-5 Query reformulations by searcher type in terms of the precision measure (two-tailed paired *t*-test)

Searcher Type	df	t-Value	p-Value	Mean of Differences
SN	54	-1.23	0.2229	-.067
DE	61	-3.02	0.0037 **	-.153
SE	52	-3.87	0.0003 ***	-.186
ML	59	-2.14	0.0368 *	-.120

Note. Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05; DE = domain experts; SN = search novices; ML = medical librarians.

4.7 User Perception of Usefulness of Displayed Terms and Search Task Difficulty

User perception of search processes like usefulness of displayed terms and search task difficulty is an important aspect of query reformulations. Previous research has suggested that the relationship between the perceived usefulness of displayed terms and search performance may depend on what kinds of users they are in an interactive search environment. User perception of search task difficulty has been shown to affect information searching behavior. In this section we look at comparisons of user perception of term usefulness and search task difficulty by searcher type and system version.

4.7.1 *Perceived usefulness of displayed terms*

User perception of term usefulness varies by type of displayed terms. More specifically, displayed terms in abstracts in MeSH– version were perceived to be more useful than displayed MeSH terms in MeSH+ version (ANOVA, $F(1, 254) = 6.19, p < .05$). Searcher type, however, did not make statistically significant differences (ANOVA, $F(3, 252) = 2.36, p > .05$).

There was a very significant interaction effect between searcher type and system version (ANOVA, $F(7, 248) = 3.11, p < .01$). These significant differences occurred in the pairs of DE-MeSH– and SE-MeSH+ (Tukey's HSD, $p < .001$) and SN-MeSH– and SE-MeSH+ (Tukey's HSD, $p < .05$) (Figure 4-12).

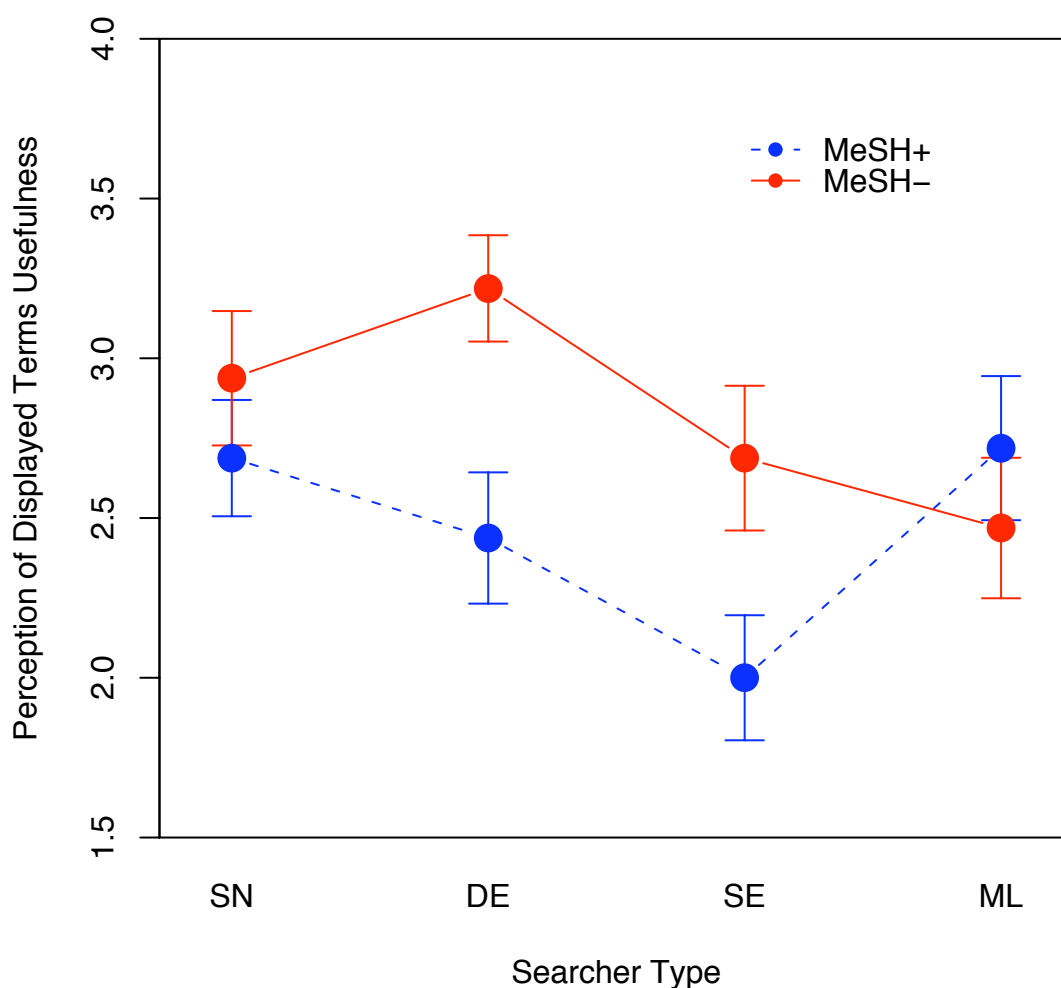


Figure 4-12 Line plot of the mean and standard error of perception of displayed terms usefulness by searcher type and system version

DEs didn't think that MeSH terms were useful, but in fact they did better in terms of the precision measure. SEs did not find MeSH terms useful relative to the abstract terms, but MLs did—this presumably reflected MLs' experience of MeSH terms. So experience does make a difference for perception of usefulness of MeSH terms. Still, SEs did not find MeSH terms useful and they did not do well in terms of the precision

measure. This is another case where perception of term usefulness is not correlated with search results.

4.7.2 Perceived search task difficulty

Searcher type made extremely significant difference in the overall ratings of perceived search task difficulty (ANOVA, $F(3, 252) = 6.13, p < .001$). SEs perceived search tasks much more difficult than DEs (Tukey's HSD, $p < .001$) and SNs (Tukey's HSD, $p < .05$). There was no difference in search task difficulty by system version (ANOVA, $F(1, 254) = 1.46, p > .05$).

Further analysis indicated that there was a strong interaction effect by searcher type and system version (ANOVA, $F(7, 248) = 2.94, p < .001$) (Figure 4-13). More specifically, SEs' searches with MeSH+ version were perceived to be more difficult than DEs' searches with either MeSH+ (Tukey's HSD, $p < .05$) or MeSH- version (Tukey's HSD, $p < .01$). Overall, SEs perceived the search task as very difficult particularly when they used MeSH+ version.

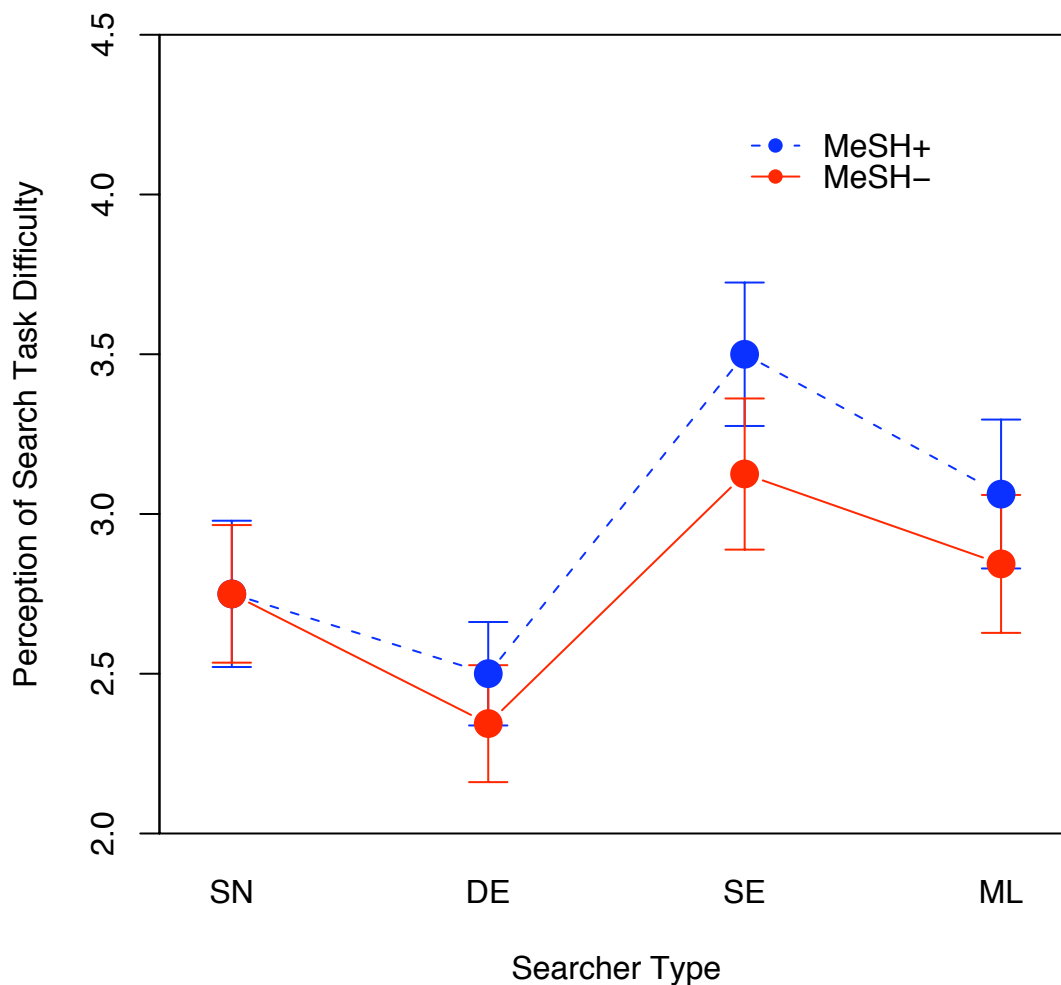


Figure 4-13 Line plot of the mean and standard error of perceived search task difficulty by searcher type and system version

Interestingly, this suggests that when SEs saw MeSH terms, it actually made them perceive the searches as more difficult. The results are consistent with the finding of correctness of comprehension test where SEs had the least domain knowledge (see Table

3-2). Because of SEs' limited domain expertise and technical nature of search topics, they considered the topics fairly difficult.

In summary, SEs did not perceive MeSH terms as useful and perceived search task as fairly difficult (Table 4-6). By contrast, DEs perceived displayed terms in abstracts as useful and did not perceive search task difficult. We speculate that user perception is correlated with domain expertise and related to relatively technical search topics. But it also might be related to lack of MeSH experience. In the next session, we will look into the question whether user perception agrees with search outcome in terms of the precision, recall and time spent measures.

Table 4-6 Summary of user perception results

	Displayed Terms Usefulness	Search Task Difficulty
System Version	MeSH- >> MeSH+	No sig. diff.
Searcher Type	No sig. diff.	SE >> DE SE >> SN
Searcher Type-	DE (MeSH-) >> SE (MeSH+)	SE (MeSH+) >> DE (MeSH+)
System Version	SN (MeSH-) >> SE (MeSH+)	SE (MeSH+) >> DE (MeSH-)

Note. >> means better at .05 level of significance; No sig. diff. = no significant difference; Sig. diff. = significant difference at .05 level; DE = domain experts; SN = search novices; ML = medical librarians; MeSH+ = MeSH+ version search system; MeSH- = MeSH- version search system.

4.8 User Perception and Outcome Measures

In this session, we will examine the relation between user perception and outcome measures. We will report the overall result and comparison by searcher type.

4.8.1 Usefulness of displayed terms and outcome measures

The overall results indicated that people's perception of whether displayed terms in abstracts helped them is positively correlated with the precision measure, but negatively correlated with the recall measure (Table 4-7). This is consistent with previous

research that shows that searcher significantly underestimates recall (e.g., Blair & Maron, 1985). More specifically, the precision obtained by searches with high score of term usefulness is by a factor 2.31, or 131% higher than searches with low score. For the recall obtained by searches with high score of term usefulness, it is by a factor of 0.36, or 64% (1-0.36) less likely to obtain higher recall. However, there was no significant relationship between usefulness of terms in abstracts and the time spent.

Table 4-7 Summary of the relation between terms in abstracts usefulness and the outcome measures (N users = 32; N questions = 20; N all searches = 256; statistical significance at 95%)

Outcome Measures	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/-	t-Value	Stat. Signif.
Precision	0.33	2.31	0.84	0.41	2.06	Yes
Recall	0.18	0.36	1.01	0.40	-2.54	Yes
Time Spent	471.55	1.52	0.42	0.37	1.13	No

The results by searcher type revealed that only DEs' perception of whether terms in abstracts helped them agrees with the precision score (Table 4-8). The precision obtained by DEs with high score of term usefulness in abstracts is by a factor of 5.32 higher than searches with low term usefulness score. The fact that DEs' perception of term usefulness is significantly correlated with precision may be due to their specialized knowledge in the biomedical domain and relatively technical search topics. This significant relation, however, was not found in other types of searchers.

Table 4-8 Summary of the relation between terms in abstracts usefulness and the precision score by searcher type (N users = 32; N questions = 20; N all searches = 128; statistical significance at 95%)

Searcher Type	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/-	t-Value	Stat. Signif.
SN	0.23	0.61	-0.49	0.73	-0.67	No
DE	0.29	5.32	1.67	0.77	2.18	Yes
SE	0.38	4.83	1.58	0.83	1.89	No
ML	0.42	0.81	-0.21	0.69	-0.31	No

Note. SN = search novices; DE = domain experts; SE = search experts; ML = medical librarians

Participants' perception of whether MeSH terms helped them does not agree with any of the outcome measures in terms of precision, recall and time spent (Table 4-9). The fact that searcher's perception of MeSH terms usefulness is not correlated with outcome measures may be attributed to the opacity of indexing terms. We speculate that searchers were unable to identify or recognize useful MeSH terms, either because they did not (1) have a chance to look at the full-text of retrieved documents, as opposed to MeSH indexers; (2) know exactly how MeSH terms are derived; or (3) have sufficient biomedical knowledge and enough time for researching the topic.

Table 4-9 Summary of the relation between MeSH term usefulness and the outcome measures (N users = 32; N questions = 20; N all searches = 256; statistical significance at 95%)

Outcome Measures	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/-	t-Value	Stat. Signif.
Precision	0.34	0.55	-0.59	0.38	-1.58	No
Recall	0.19	1.09	1.09	0.40	0.21	No
Time Spent	498.97	1.42	0.35	0.36	0.97	No

4.8.2 Perceived search task difficulty and outcome measures

The overall results indicated that people's perception of perceived search task difficulty agree with their search performance in terms of the time spent (Table 4-10). It means that the time spent by searches with high perceived search task difficulty is by a factor of 2.70, or 170% longer than searches with low difficulty. People's perception of search task difficulty, however, is not correlated with other outcome measures, including precision and recall measures. So time spent may be a measure of search task difficulty.

Table 4-10 Summary of the relation between search task difficulty and the outcome measures (N users = 32; N questions = 20; N all searches = 256; statistical significance at 95%)

Outcome Measures	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/-	t-Value	Stat. Signif.
Precision	0.34	1.01	0.01	0.26	0.05	No
Recall	0.18	0.77	-0.26	0.28	-0.94	No
Time Spent	485.26	2.70	0.99	0.26	3.78	Yes

The results by searcher type showed that only SNs' perception of search task difficulty is correlated with the time spent (Table 4-11). It means that the time spent by SNs with high perceived difficulty is by a factor of 2.98, or 198% longer than searches with low difficulty. The time spent by other types of searchers was not correlated with their perceived search task difficulty. We speculate that SNs are sensitive to perceived search task difficulty in terms of the time spent because of their background—without domain expertise and search training. SEs, who perceived the search tasks as most difficult and spent the least time searching, were not sensitive to the time spent probably because of they were not able to use their search skills on the technical topics.

Table 4-11 Summary of the relation search task difficulty and the time spent by searcher type (N users = 32; N questions = 20; N all searches = 256; statistical significance at 95%)

Searcher Type	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/-	t-Value	Stat. Signif.
SN	475.70	2.98	1.09	0.53	2.08	Yes
DE	527.48	1.28	0.25	0.54	0.46	No
SE	455.63	0.69	-0.38	0.51	-0.74	No
ML	482.23	1.17	0.16	0.51	0.30	No

Note. SN = search novices; DE = domain experts; SE = search experts; ML = medical librarians

On the whole, user perceived usefulness of terms in abstracts was correlated with the outcome measures in terms of precision and recall measures, but there was no such relation for MeSH terms. Perceived search task difficulty was correlated with the time spent. Searcher type made substantial differences in the relation between user perception and outcome measures. DEs' perception of term usefulness in abstracts was correlated with the precision measure, whereas SNs' perception of search task difficulty was correlated with the time spent.

In the next session, we will look at the correlation of user characteristics and search effectiveness in terms of the precision and recall measures.

4.9 User Characteristics

A closer examination of the relation between user characteristics and search effectiveness showed that searchers' domain knowledge, measured by the number of undergraduate/graduate level biology classes taken, was correlated with the precision measure (Table 4-12). Searchers who have taken more than five undergraduate level biology classes were estimated to obtain higher precision score by a factor of 2.57, or

157% more than those with less than five classes. Searchers who have taken more than two graduate level biology classes were by a factor of 1.82, or 82% more likely to obtain higher precision score than those with less than two courses. Other user characteristics, however, were not correlated with the precision score. These results suggested that the searcher's formal education in biology was significantly correlated with the precision of searches, probably because of the technical nature of genomics search topics.

Table 4-12 Summary of the relation between user characteristics and the precision score (N users = 32; N questions = 20; N all searches = 256; statistical significance at 95%)

User Characteristics	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/-	t-Value	Stat. Signif.
Gender	1.53	0.92	-0.09	0.26	0.26	No
Native language	1.50	1.27	0.24	0.26	0.91	No
# of undergraduate biology classes	4.94	2.57	0.94	0.29	3.21	Yes
# of graduate biology classes	1.84	1.82	0.60	0.29	2.04	Yes
# of online searching classes	3.47	1.06	0.06	0.31	0.19	No
Experience of MeSH use	0.84	0.98	-0.02	0.27	-0.09	No
Experience as information professional	0.97	1.29	0.26	0.27	0.95	No
Experience of database use	2.84	1.25	0.22	0.27	0.83	No
Frequency of database use	4.06	0.96	-0.04	0.28	-0.15	No
Age	3.59	0.77	-0.26	0.26	-0.97	No

The results from the relation between the user characteristics and the recall measure indicated that only formal search training was correlated with the recall score (Table 4-13). Searchers with more than four online searching classes were estimated to be two times more likely to obtain high recall score than searchers with fewer classes. We speculate that well-trained searchers obtained better recall score because they were able

to use more unique terms in searches than other types of searchers, and thus obtain more comprehensive results. In fact, well-trained searchers were mostly MLs who have also used the largest number of unique terms in searches (see sections 3.2.2 and 4.5.2).

However, we did not find significant correlations between other user characteristics and the recall score.

Table 4-13 Summary of the relation between user characteristics and the recall score (N users = 32; N questions = 20; N all searches = 256; statistical significance at 95%)

User Characteristics	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/-	t-Value	Stat. Signif.
Gender	1.53	1.10	0.10	0.28	0.35	No
Native language	1.50	0.82	-0.19	0.28	-0.69	No
# of undergraduate biology classes taken	4.94	0.68	-0.39	0.34	-1.16	No
# of graduate biology classes	1.84	0.53	-0.63	0.35	-1.79	No
# of online searching classes	3.47	2.00	0.69	0.32	2.18	Yes
Experience of MeSH use	0.84	1.68	0.52	0.28	1.83	No
Experience as information professional	0.97	1.48	0.39	0.29	1.36	No
Experience of database use	2.84	1.48	0.39	0.29	1.35	No
Frequency of database use	4.06	1.06	0.06	0.29	0.19	No
Age	3.59	1.33	0.29	0.28	1.04	No

Overall, domain knowledge measured by level of education was correlated with the precision score, whereas search training measured by the amount of formal training was correlated with the recall score. In the next chapter, we will discuss the role of domain knowledge and search training in search effectiveness in light of these findings.

4.10 Summary

Considering these research findings, we conclude that queries searched using MeSH and queries searched not using MeSH could achieve comparable search effectiveness in terms of the precision and recall measures (Table 14-4). The overall results suggest relatively low precision and recall score in search effectiveness and expended search effort for assigned search tasks.

Table 4-14 Summary of search effectiveness and search efficiency results

	Search Effectiveness	Search Efficiency
Overall	Mean precision = .34 Mean recall = .18	Mean = 485.3 secs/topic
System Version	MeSH+ = MeSH-	MeSH+ = MeSH-
Searcher Type	SN = DE = SE = ML	DE >> SE
Searcher Type- System Version	DE (MeSH+) >> SE (MeSH+) DE (MeSH+) >> SN (MeSH-)	DE (MeSH+) >> SE (MeSH-)

Note. >> means better at .05 level of significance; = means no significant difference; SN = search novices; DE = domain experts; SE = search experts; ML = medical librarians

We observed several significant interaction effects between searcher type and system version. When MeSH terms were offered, (1) DEs obtained better results than SEs in terms of precision, (2) DEs used more queries and query terms than SNs and (3) SEs perceived search tasks more difficult than DEs.

With regard to the relation between user perception and search performance, overall the more difficult the participant perceived search tasks to be, the more time they spent on the search tasks. If participants perceived terms in abstracts as useful, they had high precision but lower recall results.

For the relation between user characteristics and search performance, domain knowledge measured by level of education was correlated with the precision score,

whereas search training measured by the amount of formal training was correlated with the recall score.

Given these results, in the next chapter we will discuss the importance of domain knowledge and search training in the use of index terms, such as MeSH, and seen in the experimental design decisions made in this study, and reflect on the significance of this study for the assessment of the usefulness of manually assigned controlled vocabulary systems such as MeSH.

CHAPTER 5 DISCUSSION AND CONCLUSION

How useful are MeSH terms for different kinds of users? We look at the factors that affect the usefulness of MeSH terms, specifically user perceived search task difficulty, perceived usefulness of displayed MeSH terms and the correlation between user perception and the outcome measures. Our discussion also deals the role of user's domain knowledge and search training in search effectiveness and the methodological considerations of taking into account search topic variability for assessing the usefulness of MeSH+ information retrieval system in an interactive search environment. Finally, we draw conclusions from these findings and provide directions for future research.

5.1 Perceived Search Task Difficulty

Our results demonstrate that the searcher's level of domain knowledge made significant differences in perceived search task difficulty. SEs perceived the search task significantly more difficult than DEs. This is consistent with the observation that DEs had considerably higher level of domain knowledge than SEs, measured by level of formal education and correctness of comprehension test. Further, we observe that the availability of MeSH terms did not adjust participants' searching experience.

These results suggest that the intelligibility of technical search topics substantially influences searchers' perceived search task difficulty. And the search tools, such as MeSH terms, would be more useful if searchers have sufficient understanding about the search topics or the subject domain in general.

In this study searchers' perceived search task difficulty represents an inherent level of difficulty because the search tasks were chosen primarily based on the

intelligibility and required level of domain knowledge for relatively technical search topics. Users rated search tasks as difficult even though we had selected the easiest of the TREC topics, based on judgment of domain expert and medical librarian. So the TREC topics are inherently quite difficult and this may well have affected our results.

Earlier research on search task difficulty used the amount of effort measured by time as a metric (e.g., Hildreth, 2001; Wacholder & Liu, 2006; Zhang & Li, 2008). In these studies, the search tasks themselves are the source of the difficulty. Here, the inherent difficulty concerns the technical nature of genomics topics. This study provides experimental evidence that the inherent level of difficulty also affects perceived search task difficulty.

Another reason for the inherent difficulty of the TREC genomics topics is that these topics are concerned with different levels of specificity ranging from broad topics to specific gene names (Aronson et al., 2004). The sample search topic about hypertension and genes (see Figure 3-4) is a typical example of broad topics. This kind of topics is considered difficult probably because it is impossible for searchers to directly identify the gene names from search topic descriptions that may be critical for obtaining useful documents. For more specific search topics applicable to particular genes, the problem of semantic ambiguities intrinsic to biomedical terminologies makes the searching of genomics topics especially harder (see e.g. Cimino & Zhu, 2006).

The very limited time allotted for each search session also contributes to perception of search task difficulty. Given that these topics are technical in nature with varying levels of specificity and semantic ambiguities in biomedical terminologies, searchers may need more time to research the topic, formulate appropriate queries and

reformulate queries in view of search results. Searchers suggested that they might have done better if they had more time.

Finally, since we implement a Boolean-based ranked IR experimental system, users can formulate complex Boolean search expressions if they desire. The fact that formulating Boolean queries require tremendous amount of mental effort (see e.g., Cooper, 1988; Hearst & Karadi, 1997) makes the search tasks especially challenging for searchers without formal search training. Further, when searchers are assigned to the MeSH version of an experimental system, selecting the appropriate MeSH terms in view of search topics is known to be a demanding task for those who are not search experts in the biomedical domain (e.g., Lowe & Barnett, 1994; Nelson, Johnston & Humphreys, 2001).

In spite of the inherent difficulty of the topics, participants did not discontinue assigned search tasks during the experiment. In particular the SEs and MLs might have done better with search tasks that were more like what they were used to.

Readers are advised to interpret the study findings considering the experimental design within an academic research environment—motivated and observed users conducting searches of relatively difficult tasks within time constraints. Besides the factors that may contribute to perceived search task difficulty, our next discussion is concerned with the relation between the perceived search task difficulty and the outcome measures in terms of precision, recall and time spent.

5.2 Perceived Search Task Difficulty and Outcome Measures

Overall the more difficult the participant perceived search tasks to be, the more time they spent on the search tasks. However, this correlation was only significant for

SNs when we considered differences in searcher type.

Our results also show that overall the time spent is a good indicator of perceived search task difficulty. Previous research has indicated that user perceived search task difficulty was correlated with the time spent for factual search tasks in Web searches (Gwizdka & Spence, 2006; Kim, 2008). Our results further demonstrate that this correlation varies by searcher type in ad hoc search tasks using a Boolean-based information retrieval system. This source of variation may come from searcher differences in terms of level of domain knowledge and search training. Taken together, perceived search task difficulty is a function of the individual differences, especially level of domain knowledge and search training, and also of types of search tasks.

In spite of the overall non-significant correlation between perceived search task difficulty and search performance in terms of precision measure, our findings reveal a significant correlation under specific conditions. When MeSH terms were offered, DEs and SEs' perceived search task difficulty agreed with their search performance in terms of precision measure. This suggests that to help users obtain better search results using search tools like MeSH terms, user perceived search task difficulty could be useful for personalizing search results if we can infer the searcher's characteristics of level of domain knowledge and search training from measures of search behaviors and perceived usefulness of displayed terms.

5.3 Usefulness of Displayed Terms and Search Effectiveness

The present study constitutes another illustration of the usefulness of displayed terms in abstracts in an interactive information retrieval environment. Our overall results indicate that if participants perceived terms in abstracts as useful, they had higher

precision but lower recall results. It is likely that they are able to recognize search terms that are critical for search topics during search processes.

For participants who saw the MeSH+ system (i.e. abstracts and MeSH terms), there was no correlation between their perception of usefulness of the MeSH terms and their performance as measured by precision and recall. Unfortunately, we only asked users of the MeSH+ system about the usefulness of the MeSH terms. We don't know exactly what the results might be if abstracts were removed from MeSH+ system.

The fact that DEs' perception of usefulness of terms in abstracts was significantly correlated with precision score may be due to their specialized knowledge in the biomedical domain. Interestingly, DEs obtained significantly better precision score than SEs using MeSH+ system version, but they did not perceive MeSH terms as very useful. This is a case where user's perception of displayed terms usefulness and search performance depends on the available search tool. It is also a nice example of discrepancy between user perception and search performance.

In this study the difficulty of topics may be responsible for non-significant relation between displayed terms usefulness and search effectiveness by SEs and MLs and reduced performance. Our findings indirectly support previous studies that compared the search performance among different types of searchers, in which medical librarians can be better than clinical end-users (physicians, physician trainees and clinicians) in terms of precision and recall measures in searching clinical topics (Haynes, McKibbin, Walker, Ryan, Fitzgerald, & Ramsden, 1990; Hersh & Hickam, 1994; McKibbin et al., 1990). Because the medical librarians had the knowledge to understand the clinical topics and made good use of system features, such as displayed MeSH terms, their search

results were better than other clinical end-users.

The significant relation between perceived usefulness of terms in abstracts and precision score by DEs, and the fact that they were able to improve search results by query reformulations imply that identification and extraction of terms from abstracts and displaying those terms in an organized way (Wacholder & Liu, 2006, 2008), might be useful for searchers engaged in interactive query reformulation and ultimately obtaining better search results.

The disagreement between perceived usefulness of MeSH terms and search performance by DEs suggests that MeSH terms would be more useful for searchers with domain expertise if they were presented to searchers in alternative ways like grouping search results by MeSH terms, or some kind of direct system interventions would be necessary. On the other hand, controlled vocabulary like MeSH terms would not be useful for searchers without substantial domain expertise, such as SEs in our study, if they were only presented to searchers in displayed bibliographic records.

Overall, these findings advance our understanding of the usefulness of displayed terms for different kinds of searchers and the specific conditions under which searchers' perceived usefulness of displayed terms agree or disagree with their search effectiveness.

5.4 Domain knowledge

This study indicates that domain knowledge plays an important role in effective use of MeSH terms, especially when the search topics are technical in nature.

Specifically, MeSH terms are most useful in terms of precision for domain experts. Our results therefore contradict earlier research (e.g., Allen, 1991; Pao et al., 1993) that has suggested that domain knowledge is not correlated with search outcome. This may be

because the earlier research used a small number of search topics and homogenous group of participants with relatively similar subject background. Our study has demonstrated that the use of a relatively large number of search topics in an interactive search environment through experimental design is feasible. Our participant's considerable differences in level of domain knowledge have made it possible to observe the subtle differences in search performance.

One prominent finding from this study is the conditions in which searcher's domain knowledge makes a difference in search performance. More specifically, we identified significant interactional effects between levels of domain knowledge and system versions in terms of the precision measure. It suggests that searchers can benefit the most from the proper use of search tools, such as MeSH terms, when they have sufficient knowledge about the search topic. That is, the interactional effects between the searcher's high level of domain knowledge and the use of MeSH terms contribute to significantly better precision scores. However, users obtained better recall score when they searched in a best match IR system using thesaurus-based query enhancement features (Jones et al., 2005). Other naturalistic-oriented studies (e.g., Nielsen, 2004; Sihvonen & Vakkari, 2004) have indicated that domain expert searchers can benefit more than search novices from the use of thesaurus tools in selecting potentially useful terms for expanding initial queries. Overall, this study provides empirical evidence that controlled vocabularies are useful for obtaining better precision score when searchers have substantial knowledge about the topic using a Boolean-based IR system with ranked functions. However, we did not find evidence that these terms are useful for other kinds of users.

The differences in the level of domain knowledge of the search topic being searched revealed in the number of terms and queries used per search session by type of user. This finding is consistent with prior research on end-user searching behaviors in view of scientific domain topics in controlled experiments (e.g., Allen, 1991; Wildemuth, 2004), indicating that high-knowledge searchers tend to use more search expressions when they search relatively difficult search topics.

Since query reformulation tasks are concerned with forming mental representations of the search topic, and translating those into search expressions, searcher's domain knowledge and verbal ability are crucial for the execution of these tasks. There is some evidence that searcher's verbal ability was associated with search performance in controlled user experiments (e.g., Bellardo, 1985; Dumais & Schmitt, 1991; Saracevic & Kantor, 1988). Verbal ability is a user characteristic that deserves further investigation.

Our findings suggest that search novices do not know how to improve search results by query reformulations when they search technical topics. However, the domain experts in the present study were all non-native speakers of English with high-level domain knowledge from diverse subfields of biology and they achieved the best search performance among the groups. So we conclude that domain knowledge is more important than language skills in searches on technical search topics.

5.5 Search Training

The results demonstrate that the level of formal search training was correlated with search outcome. More specifically, searchers with more than four online searching classes were estimated to be two times more likely to obtain high recall score than

searchers with fewer classes. Our results therefore contradict prior research (e.g., Fenichel, 1981; McKibbin et al., 1990; Howard, 1982; Pao et al., 1993) that has suggested that search experience as a theoretical construct was not correlated with the recall score.

One possible explanation is that earlier studies depended on the user's subjective self-reporting data to measure the search experience rather than objective criteria. Some researchers have reservations about the reliability of using subjective self-reporting of exposure to IR systems in the measurement of search experience (e.g., Dalrymple & Zweizig, 1992; Moore et al., 2007). This study primarily used the objective criterion of formal search training, supplemented by self-reporting of frequency and years of experience searching online databases. We found that the number of online searching classes was a good predictor of recall score, while the frequency and years of use experience were not. We speculate that searchers with high level of search training were able to identify more potentially useful terms in query formulation contributing to significantly better search performance in terms of recall.

5.6 Search Topic Variability

Our experimental design of using relatively large number of search topics represents an initial step to address the variability of search topics within an interactive IR experiment. We made the assumption that all the search topics were equally difficult so that a total of twenty search topics, or ten search topic pairs, can be used and controlled with searcher types and system versions by a Graeco-Latin squared design. With carefully selected search topics and experimental design, we were able to detect subtle differences in search performance by different types of searchers interacting with

an experimental system with/without MeSH terms.

Since the search topic pairs were controlled with searcher types and system versions, this allowed us to determine the impact of search topics on search performance. The use of a large number of search topics increased the external validity in terms of the possible queries received by the IR system. However, for the current study search topic variability remained the most significant factor that affects search performance.

To tackle the problem of search topic variability, researchers have used a large number of search topics in standard ad hoc retrieval tasks in TREC activities (Buckley & Voorhees, 2005). Because the primary purpose was to compare the search performance of different retrieval techniques, the researcher can be confident about the relative system ranking by averaging the performance scores across the topics. Using a large number of topics (e.g., 50 topics in TREC standard ad hoc task) in an IR experiment involving human searchers was not feasible due to limited resources. Other proposals, such as the use of artificial topics for directly controlling the variables coupled with topics (Robertson, 1981), have not been implemented partly because we have limited understanding of the nature and characteristics of search topics, and do not offer a good typology of search topics for evaluation purposes. So we used balanced experimental design to consider search topic variability within an interactive search environment.

Methodologically, this study demonstrates a user experiment design that can be used to investigate the impact of specific user characteristics on search performance. One primary feature was the control of search topic pairs, searcher types and the two system versions being compared. This allowed us to assess the effects of user characteristics and system differences, with the use of relatively large number of search topics.

This approach has advantages over the commonly used matched-pair design in which the same set of topics was used to the systems being compared. It is also not a non-matched-pair design, as discussed in Robertson (1990), whereby the independent sample method was used to obtain search topics.

One example described in Robertson, Thompson and Macaskill (1986) required a very large sample of independent searches (e.g., 500 topics). A relatively small number of search topics was usually used in a matched-pair design study (between two and eight), while a non-matched-pair design required a very large number of search topics. Another related experimental design option that a list of relatively large number of search topics (e.g., 24 topics) was prepared and presented to searchers in random order with time limit for each topic, can produce robust statistical results using enough subject and topic samples, but the frequency distribution of the times each topic was searched may not be evenly distributed (e.g., Wacholder & Liu, 2006, 2008). Our method required many fewer topics and participants.

5.7 Evaluation of MeSH+ and MeSH– Information Retrieval Systems

Our results provide experimental evidence of the usefulness of MeSH+ information retrieval systems in an interactive search environment. Previous research has compared the retrieval performance of MeSH+ vs. automatic indexing techniques in laboratory settings without the involvement of end-users of IR systems (e.g., Salton, 1969, 1972; Savoy, 2005). One general conclusion has been that the retrieval performance obtained by automatic indexing with various combinations of retrieval models can be as effective as that obtained by MeSH+ system.

In the present study, MeSH terms were useful for domain experts' searches on

technical topics but we did not find evidence that MeSH terms were useful for other kinds of searchers. Although domain experts did not have experiences using MeSH terms before participating our study, they were able to benefit from the use of MeSH terms and obtain better search results in terms of precision. Interestingly, domain experts spent the most time and issued the most queries, but they did not perceive that MeSH terms were useful for their searches. Our results therefore advance our understanding of the impact of user factors on search results using controlled vocabularies in an interactive search environment.

The experimental results strongly suggest that searchers with substantial domain knowledge can benefit from the use of MeSH terms in terms of the precision measure, even though their perception of the usefulness of MeSH terms did not agree with search performance. The previous finding that retrieval performance obtained by automatic indexing with various combinations of search models can be as effective as that obtained by manual indexing in a batch mode evaluation (e.g., Abdou & Savoy, 2008), therefore can be extended to the context of controlled user experiment. Our results provide another explanation for the transferability of IR system improvement to user search performance in which the user characteristic of domain knowledge plays an important role (cf. Hersh, Turpin, Price, Chan, Kraemer, Sacherek, L., et al. 2000; Turpin & Scholer, 2006; Turpin & Hersh, 2001).

The results from a comparison between the machine runs in TREC Genomics track 2004 and the human searches in the present study indicated very high correlations in terms of MAP, P10 and P100 measures by search topic. In other words, the same topics were difficult for human and machine runs. It suggests that the search

improvement in automatic IR systems can also benefit human searches.

It also suggests that the enhancement in retrieval techniques for poorly performed topics may also benefit human searches. For human searchers without advanced knowledge in biology, it would be hard to specify queries when searching this kind of topics. More detailed analyses of failures with respect to the poorly performed topics by human searchers or machine runs, such as the qualitative analysis employed in previous research (e.g., IJzereef et al., 2005; Lancaster, 1969; Savoy, 2007), may contribute to our understanding of the nature and properties of search topics and to the improvement in the consistency of retrieval technology.

Methodologically, this study demonstrates the feasibility and reuse of using a test collection, originally created for evaluating the effectiveness of retrieval techniques in a traditional ad hoc task, for a controlled user experiment. A relatively large-scale user experiment similar to this one would not be possible without a well-constructed test collection, including a large set of test documents, search topics and relevance judgments. We carefully considered the completeness of test documents, intelligibility of search topics, and reliability of relevance judgments for a user experiment. Similar to many IR experiments, we also dealt with a test collection with incomplete relevance judgments. The analysis of pooled relevance judgment set shows that reused data was reliable for this experiment, even though we did not contribute retrieved documents to the original judgment pool. But the limitation is that the topics were very hard for human searchers.

5.8 Conclusion

This study was designed to determine how useful MeSH terms are for different kinds of users. We observed four different kinds of information seekers using an

experimental information retrieval system: 1) search novices; 2) domain experts; 3) search experts and 4) medical librarians. The search tasks were a subset of the relatively difficult topics originally created for Text REtrieval Conference (TREC) Genomics track. Effectiveness of retrieval was based on the relevance judgments provided by TREC. Participants searched either using a version of the system in which abstracts and MeSH terms were displayed or another version in which they had to formulate their own terms based only on the display of abstracts.

Our results provide experimental evidence of the usefulness of MeSH terms in an interactive search environment. Previous research has compared the retrieval performance of MeSH terms and automatic indexing techniques in laboratory settings without human searchers. Salton (1972) claimed that “*fully automatic text processing methods can be used to obtain retrieval output of an effectiveness substantially equivalent to that provided by conventional, manual indexing* (emphasis original) (p. 81).” Our results support this general conclusion and further identify the factors that have significant impact on the search effectiveness of MeSH terms.

The experimental results suggest that searchers with substantial domain knowledge can benefit from the use of MeSH terms in terms of the precision measure, even though domain experts did not perceive that MeSH terms were useful. Levels of domain knowledge were reflected in domain experts’ use of more search expressions, more issued queries and more time spent than other types of searchers using MeSH+ system. Domain experts’ perception of the usefulness of terms in abstracts agrees with their precision score. We found that domain knowledge is more important than language skills in searches on technical topics.

Search experts did not find MeSH terms useful and they did not do well in searches. Search experts perceived the search task as very difficult especially when they used MeSH+ system. The results from comprehension test revealed that the technical topics are especially challenging for search experts. It suggests that search training alone cannot compensate for the lack of domain knowledge in searches on technical topics as we used in the study.

Medical librarians had the most search training in our study, but their knowledge in biology was not as high as we expected. Searchers' level of formal search training was a good predictor of search effectiveness in terms of the recall measure. The impact of search training was also reflected in the use of MeSH terms and the number of unique terms per search session: the more search training one had, the more likely one would use MeSH terms. Medical librarians used significantly more unique terms than search novices when MeSH+ system was offered. MeSH terms, originally designed for medical librarians in intermediary searching, were not very helpful in our assigned search tasks.

Searchers with low level of search training, especially search novices can search reasonably well in part because of the state-of-the-art retrieval system used in this study. Even though search novices expended the least effort in terms of the number of search terms and issued queries, Search novices' use of MeSH+ system improved their search results in terms of precision. However, search novices did not know how to improve search results in terms of precision in query reformulations.

The high correlation between human and computer search results by search topic suggests that improvement in retrieval algorithms also benefits human searches. The better search performance achieved by computer searches maybe due to the fact that

these systems are specially trained for searches on technical genomics topics and these topics were originally designed for experimental purposes of comparing the search effectiveness of different retrieval techniques.

Methodologically, this study has demonstrated the feasibility and reusability of using a test collection, originally created for evaluating the effectiveness of retrieval techniques in an ad hoc search task, for a controlled user experiment. We used a relatively large number of search topics in a user experiment through experimental design techniques. The reliability of relevance judgment sets was ensured by additional analysis of top 10 search results from our human searches. By the experimental design and methodology, we were able to detect the subtle differences in search effectiveness obtained by different kinds of users.

5.9 Future Research

The experimental design and methodology similar to our study can be used to assess the quality of automatically extracted phrases as displayed index terms in support of browsing or interactive query expansion tasks. For example, recently researchers have proposed several methods of automatic identification of index terms to support interactive information retrieval tasks (e.g., Anick & Tipirneni, 1999; Edgar et al., 2003; Wacholder, Evans, & Klavans, 2001). To make these search tools more useful for end-users in operational systems, it is especially important to assess the impact of displayed index terms and searchers characteristics on search effectiveness.

As mentioned earlier, we used genomics search topics that were technical and difficult for most searchers in our experiment. We decided to use these difficult search topics because of the availability of the TREC relevance judgments, and because it allows

us to compare the usefulness of human created terms to standard retrieval techniques. It is recommended to use other search topics, such as clinical topics, for assessing the usefulness of MeSH terms in the settings of medical or hospital libraries in future research. The use of clinical topics might have different results for medical librarians because they are more familiar with these topics. This will also enhance the external validity of this kind of study since medical librarians usually receive clinical topics in their work settings.

Our study provides empirical evidence that the effort to create MeSH terms is worthwhile for domain experts' searches on technical topics. This study has focused on the impact of displayed controlled index terms on search effectiveness. For comparative purposes we used MeSH terms as a case study of the impact of controlled vocabularies and used a Boolean-based retrieval system with ranking functions. We only compared one kind of controlled vocabulary within a single IR system because we were concerned with the difficulty of separating out the effects of users, systems and topics in an interactive search environment. It is recommended to replicate this study using different kind of controlled vocabulary and search topics before we generalize the results to other settings.

Bibliography

- Abdou, S., & Savoy, J. (2008). Searching in MEDLINE: Query expansion and manual indexing evaluation. *Information Processing & Management*, 44(2), 781-789.
- Allen, B. (1991). Topic knowledge and online catalog search formulation. *Library Quarterly*, 61(2), 188-213.
- Anderson, J. D., & Pérez-Carballo, J. (2001). The nature of indexing: How humans and machines analyze messages and texts for retrieval. I. Research and the nature of human indexing. *Information Processing & Management*, 37(2), 231-54.
- Anderson, J. D., & Pérez-Carballo, J. (2005). *Information retrieval design: Principles and options for information description, organization, display, and access in information retrieval databases, digital libraries, catalogs, and indexes*. St. Petersburg, FL: Ometeca Institute.
- Anick, P. G., & Tipirneni, S. (1999). The paraphrase search assistant: Terminological feedback for iterative information seeking. *Proceedings of the ACM SIGIR Conference*, 22, 153-9.
- Aronson, A. R., Demner, D., Humphrey, S. M., Ide, N. C., Kim, W., Liu, H., et al. (2004). Knowledge-intensive and statistical approaches to the retrieval and annotation of genomics MEDLINE citations. In E. M. Voorhees & L. P. Buckland (Eds.), *The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)*. Retrieved April 11, 2009, from <http://trec.nist.gov/pubs/trec13/papers/nlm-umd-ul.geo.pdf>
- Aronson, A. R., Bodenreider, O., Chang, H. F., Humphrey, S. M., Mork, J. G., Nelson, S. J., et al. (2000). The NLM indexing initiative. In J. M. Overhage (Ed.), *Proceedings of AMIA Symposium*, 17-21. Retrieved April 11, 2009, from <http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=2243970&blobtype=pdf>
- Austin, D. (1976). PRECIS in a multilingual context: Part 1. PRECIS: An overview. *Libri*, 26(1), 1-37.
- Banks, D., Over, P., & Zhang, N. (1999). Blind men and elephants: Six approaches to TREC data. *Information Retrieval*, 1(1/2), 7-34.
- Barry, C. L. (2005). Establishing a research agenda for online search behaviors. *Proceedings of the ASIS&T Annual Meeting*, 42. Retrieved April 11, 2009, from <http://www3.interscience.wiley.com/cgi-bin/fulltext/112785680/PDFSTART>
- Bean, C. A., & Green, R. (Eds.). *Relationships in the organization of knowledge*. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Beaulieu, M., Robertson, S., & Rasmussen, E. (1996). Evaluating interactive systems in

- TREC. *Journal of the American Society for Information Science*, 47(1), 85-94.
- Belkin, N. J., Cool, C., Head, J., Jeng, J., Kelly, D., Lin, S., et al. (2000). Relevance feedback versus local context analysis as term suggestion devices: Rutgers' TREC-8 interactive track experience. In E. M. Voorhees & D. K. Harman (Eds.), *The Eighth Text REtrieval Conference (TREC-8)*, 565-573. Retrieved April 11, 2009, from <http://trec.nist.gov/pubs/trec8/papers/ruint.pdf>
- Belkin, N. J., Kelly, D., Kim, G., Kim, J. Y., Lee, H. J., Muresan, G., et al. (2003). Query length in interactive information retrieval. *Proceedings of the ACM SIGIR Conference*, 26, 205-212.
- Bellardo, T. (1985). An investigation of online searcher traits and their relationship to search outcome. *Journal of the American Society for Information Science*, 36(4), 241-250.
- Bibliographic Services Task Force. (2005). *Rethinking how we provide bibliographic services for the University of California*. Retrieved September 17, 2006, from <http://libraries.universityofcalifornia.edu/sopag/BSTF/Final.pdf>
- Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3), 289-99.
- Buckley, C. (1999). trec_eval IR evaluation package [Computer software]. Retrieved November 16, 2007, from <ftp://ftp.cs.cornell.edu/pub/smart/>
- Buckley, C., & Voorhees, E. M. (2005). Retrieval system evaluation. In E. M. Voorhees, & D. K. Harman (Eds.), *TREC: Experiment and evaluation in information retrieval* (pp. 53-75). Cambridge, MA: The MIT Press.
- Buckley, C., Dimmick, D., Soboroff, I., & Voorhees, E. (2007). Bias and the limits of pooling for large collections. *Information Retrieval*, 10(6), 491-508.
- Calhoun, K. (2006). The changing nature of the catalog and its integration with other discovery tools. Retrieved August 14, 2006, from <http://www.loc.gov/catdir/calhoun-report-final.pdf>
- Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: R. McNally.
- Champely, S. (2007). Pwr: Basic functions for power analysis (R package version 1.1) [Computer software]. Retrieved April 11, 2009, from <http://cran.r-project.org/web/packages/pwr/index.html>
- Cimino, J. J., & Zhu, X. (2006). The practical impact of ontologies on biomedical informatics. *Yearbook of Medical Informatics*, 124-135.
- Cleverdon, C. W. (1967). The Cranfield tests on index language devices. *Aslib*

- Proceedings*, 19(6), 173-193.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates.
- Cooper, W. S. (1988). Getting beyond boole. *Information Processing and Management*, 24(3), 243-248.
- Croft, W. B. (1989). Automatic indexing. In B. H. Weinberg (Ed.), *Indexing: The state of our knowledge and the state of our ignorance* (pp. 86-100). Medford, NJ: Learned Information.
- Dalrymple, P. W., & Zweizig, D. L. (1992). Users' experience of information retrieval systems: An exploration of the relationship between search experience and affective measures. *Library & Information Science Research*, 14(2), 167-81.
- Dextre Clarke, S. G. (2008). The last 50 years of knowledge organization: A journey through my personal archives. *Journal of Information Science*, 34(4), 427-437.
- Drabenstott, K. M., Simcox, S., & Williams, M. (1999). Do librarians understand the subject headings in library catalogs? *Reference and User Services Quarterly*, 38(4), 369-87.
- Dumais, S. T., & Belkin, N. J. (2005). The TREC interactive track: Putting the user into search. In E. M. Voorhees, & D. K. Harman (Eds.), *TREC: Experiment and evaluation in information retrieval* (pp. 123-152). Cambridge, MA: The MIT Press.
- Dumais, S. T., & Schmitt, D. G. (1991). Iterative searching in an online database. In *Proceedings of the Human Factors Society 35th Annual Meeting* (pp. 398-402). Santa Monica, CA: Human Factors Society.
- Edgar, K. D., Nichols, D. M., Paynter, G. W., Thomson, K., & Witten, I. H. (2003). A user evaluation of hierarchical phrase browsing. *Lecture Notes in Computer Science*, 2769, 313-324.
- Efthimiadis, E. N. (1996). Query expansion. *Annual Review of Information Science and Technology*, 31, 121-187.
- Efthimiadis, E. N. (2000). Interactive query expansion: A user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science*, 51(11), 989-1003.
- Thesaurus of engineering and scientific terms. (1967). Washington, DC: Department of Defense.
- Fenichel, C. H. (1981). Online searching: Measures that discriminate among users with different types of experiences. *Journal of the American Society for Information Science*, 32(1), 23-32.

- Fidel, R. (1991). Searchers' selection of search keys: II. Controlled vocabulary or free-text searching. *Journal of the American Society for Information Science*, 42(7), 501-514.
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.
- Fleiss, J. L., Levin, B. A., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: John Wiley.
- Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.
- Gross, T., & Taylor, A. G. (2005). What have we got to lose? The effect of controlled vocabulary on keyword searching results. *College & Research Libraries*, 66(3), 212-230.
- Gwizdka, J., & Spence, I. (2006). What can searching behavior tell us about the difficulty of information tasks? A study of Web navigation. *Proceedings of the American Society for Information Science and Technology*, 43. Retrieved April 11, 2009, from <http://www3.interscience.wiley.com/cgi-bin/fulltext/116328847/PDFSTART>
- Harman, D. K. (1993). Overview of the first Text REtrieval Conference (TREC-1). In D. K. Harman (Ed.), *The First Text REtrieval Conference (TREC-1)*, 1-20. Retrieved April 11, 2009, from <http://trec.nist.gov/pubs/trec1/papers/01.txt>
- Haynes, R. B., McKibbin, K. A., Walker, C. J., Ryan, N., Fitzgerald, D., & Ramsden, M. F. (1990). Online access to MEDLINE in clinical settings: A study of use and usefulness. *Annals of Internal Medicine*, 112(1), 78-84.
- Hearst, M. A., & Karadi, C. (1997). Cat-a-cone: An interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. *Proceedings of the ACM SIGIR Conference*, 31, 246-255.
- Hembrooke, H. A., Granka, L. A., Gay, G. K., & Liddy, E. D. (2005). The effects of expertise and feedback on search term selection and subsequent learning. *Journal of the American Society for Information Science and Technology*, 56(8), 861-871.
- Hersh, W., Buckley, C., Leone, T. J., & Hickam, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. *Proceedings of the ACM SIGIR Conference*, 17, 192-201.
- Hersh, W., & Hickam, D. (1994). Use of a multi-application computer workstation in a clinical setting. *Bulletin of the Medical Library Association*, 82(4), 382-389.
- Hersh, W., Turpin, A., Price, S., Chan, B., Kraemer, D., Sacherek, L., et al. (2000). Do batch and user evaluations give the same results? *Proceedings of the ACM SIGIR Conference*, 23, 17-24.

- Hersh, W. R., Bhuptiraju, R. T., Ross, L., Johnson, P., Cohen, A. M., & Kraemer, D. F. (2004). TREC 2004 genomics track overview. In E. M. Voorhees & L. P. Buckland (Eds.), *The Thirteenth Text REtrieval Conference (TREC-13)*. Retrieved November 16, 2007, from <http://trec.nist.gov/pubs/trec13/papers/GEO.OVERVIEW.pdf>
- Hersh, W. R. (2008). *Information retrieval: A health and biomedical perspective* (3rd ed.). New York: Springer.
- Hersh, W., Bhupatiraju, R., Ross, L., Roberts, P., Cohen, A., & Kraemer, D. (2006). Enhancing access to the Bibliome: The TREC 2004 genomics track. *Journal of Biomedical Discovery and Collaboration*, 1(1). Retrieved November 16, 2007, from <http://www.j-biomed-discovery.com/content/1/1/3>
- Hildreth, C. R. (2001). Accounting for users' inflated assessments of on-line catalogue search performance and usefulness: An experimental study. *Information Research*, 6(2). Retrieved November 16, 2007, from <http://informationr.net/ir/6-2/paper101.html>
- Howard, H. (1982). Measures that discriminate among online searchers with different training and experience. *Online Review*, 6(4), 315-327.
- Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science*, 44(3), 161-174.
- Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. *Proceedings of the ACM SIGIR Conference*, 16, 329-338.
- Humphrey, S. M., Rogers, W. J., Kilicoglu, H., Demner-Fushman, D., & Rindfleisch, T. C. (2006). Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology*, 57(1), 96-113.
- Ijzereef, L., Kamps, J., & de Rijke, M. (2005). Biomedical retrieval: How can a thesaurus help? *Lecture Notes in Computer Science*, 3761, 1432-1448.
- Joho, H., Sanderson, M., & Beaulieu, M. (2004). Study of user interaction with a concept-based interactive query expansion support tool. *Lecture Notes in Computer Science*, 2997, 42-56.
- Jones, S., Gatford, M., Robertson, S., Hancock-Beaulieu, M., Secker, J., & Walker, S. (1995). Interactive thesaurus navigation: Intelligence rules OK? *Journal of the American Society for Information Science*, 46(1), 52-59.
- Keen, E. M. (1973). The Aberystwyth index languages test. *Journal of Documentation*, 29(1), 1-35.
- Keen, E. M. (1977). On the generation and searching of entries in printed subject indexes.

Journal of Documentation, 33(1), 15-45.

- Kim, J. (2008). Perceived difficulty as a determinant of web search performance. *Information Research*, 13(4). Retrieved January 8, 2009, from <http://InformationR.net/ir/13-4/paper379.html>
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.
- Lagergren, E., & Over, P. (1998). Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment. *Proceedings of the ACM SIGIR Conference*, 21, 164-72.
- Lagoze, C., Krafft, D., Cornwell, T., Dushay, N., Eckstrom, D., & Saylor, J. (2006). Metadata aggregation and "automated digital libraries": A retrospective on the NSDL experience. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 6, 230-239.
- Lancaster, F. W. (1969). MEDLARS: Report on the evaluation of its operating efficiency. *American Documentation*, 20(2), 119-142.
- Larson, R. R. (1991). The decline of subject searching: Long-term trends and patterns of index use in an online catalog. *Journal of the American Society for Information Science*, 42(3), 197-215.
- Lesk, M. E., & Salton, G. (1968). Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4(4), 343-359.
- Lesk, M. (2004). *Understanding digital libraries* (2nd ed.). San Francisco: Morgan Kaufmann.
- Lowe, H. J., & Barnett, G. O. (1994). Understanding and using the Medical Subject Headings (MeSH) vocabulary to perform literature searches. *JAMA: The Journal of the American Medical Association*, 271(14), 1103-1108.
- Lu, Z., Kim, W., & Wilbur, W. (2009). Evaluation of query expansion using MeSH in PubMed. *Information Retrieval*, 12(1), 69-80.
- Marchionini, G., & Dwiggins, S. (1990). Efforts of search and subject expertise on information seeking in a hypertext environment. *Proceedings of the ASIS Annual Meeting*, 27, 129-142.
- Markey, K. (2007a). Twenty-five years of end-user searching, part 1: Research findings. *Journal of the American Society for Information Science and Technology*, 58(8), 1071-81.
- Markey, K. (2007b). Twenty-five years of end-user searching, part 2: Future research directions. *Journal of the American Society for Information Science and*

Technology, 58(8), 1123-30.

McKibbin, K. A., Haynes, R. B., Walker Dilks, C. J., Ramsden, M. F., Ryan, N. C., Baker, L., et al. (1990). How good are clinical MEDLINE searches? A comparative study of clinical end-user and librarian searches. *Computers and Biomedical Research*, 23(6), 583-593.

Meadow, C. T., Marchionini, G., & Cherry, J. M. (1994). Speculations on the measurement and use of user characteristics in information retrieval experimentation. *Canadian Journal of Information and Library Science*, 19(4), 1-22.

Meadow, C. T., Wang, J., & Yuan, W. (1995). A study of user performance and attitudes with information retrieval interfaces. *Journal of the American Society for Information Science*, 46(7), 490-505.

MeSH Browser (2003 MeSH) (2004). Retrieved November 16, 2007, from <http://www.nlm.nih.gov/mesh/2003/MBrowser.html>

Moore, J. L., Erdelez, S., & Wu, H. (2007). The search experience variable in information behavior research. *Journal of the American Society for Information Science and Technology*, 58(10), 1529-1546.

Morales, M. (2009). Sciplot: Scientific graphing functions for factorial designs (R package version 1.0-4) [Computer software]. Retrieved April 11, 2009, from <http://cran.r-project.org/web/packages/sciplot/index.html>

National Library of Medicine (U.S.). (1960). Medical Subject Headings.

Nelson, S. J., Johnston, W. D., & Humphreys, B. L. (2001). Relationships in medical subject headings (MeSH). In C. A. Bean, & R. Green (Eds.), *Relationships in the organization of knowledge* (pp. 171-184). Dordrecht, The Netherlands: Kluwer Academic Publishers.

New Zealand Digital Library Project. (2006). Greenstone digital library software (version 2.70) [computer software]. Hamilton, New Zealand: Department of Computer Science, The University of Waikato.

Nielsen, M. L. (2004). Task-based evaluation of associative thesaurus in real-life environment. *Proceedings of the ASIS&T Annual Meeting*, 41, 437-447.

Nitecki, D. A., & Abels, E. (Eds.). *Advances in librarianship*. London: Academic Press.

On the record: Report of the library of congress working group on the future of bibliographic control (2008). Retrieved April 11, 2009, from <http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf>

Palmquist, R. A., & Kim, K. S. (2000). Cognitive style and on-line database search

- experience as predictors of web search performance. *Journal of the American Society for Information Science*, 51(6), 558-566.
- Pao, M. L., Grefsheim, S. F., Barclay, M. L., Woolliscroft, J. O., McQuillan, M., & Shipman, B. L. (1993). Factors affecting students use of MEDLINE. *Computers and Biomedical Research*, 26(6), 541-555.
- Paynter, G. W. (2005). Developing practical automatic metadata assignment and evaluation tools for internet resources. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 5, 291-300.
- PubMed searches. (2007). Retrieved October 1, 2008, from http://www.ncbi.nlm.nih.gov/About/tools/restable_stat_pubmeddata.html
- R Development Core Team. (2008). R: A language and environment for statistical computing (version 2.8.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved April 11, 2009, from <http://lib.stat.cmu.edu/R/CRAN/>
- Robertson, S. E. (1981). The methodology of information retrieval experiment. In K. Sparck Jones (Ed.), *Information retrieval experiment* (pp. 9-31). London: Butterworth.
- Robertson, S. E. (1990). On sample sizes for non-matched-pair IR experiments. *Information Processing and Management*, 26(6), 739-53.
- Robertson, S. E., Thompson, C. L., & Macaskill, M. J. (1986). Weighting, ranking and relevance feedback in a front-end system. *Journal of Information and Image Management*, 12(1/2), 71-75.
- Rowley, J. (1994). The controlled versus natural indexing languages debate revisited: A perspective on information retrieval practice and research. *Journal of Information Science*, 20(2), 108-119.
- Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. *Proceedings of the ACM SIGIR Conference*, 26, 213-220.
- Salton, G. (1969). A comparison between manual and automatic indexing methods. *American Documentation*, 20(1), 61-71.
- Salton, G. (1972). A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of the American Society for Information Sciences*, 23(2), 75-84.
- Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM*, 29(7), 648-56.
- Saracevic, T. (1991). Individual differences in organizing, searching and retrieval

- information. *Proceedings of the ASIS Annual Meeting*, 28, 82-86.
- Saracevic, T., & Kantor, P. (1988). A study of information seeking and retrieving. II. users, questions, and effectiveness. *Journal of the American Society for Information Science*, 39(3), 177-96.
- Saracevic, T., Kantor, P., Chamis, A. Y., & Trivison, D. (1988). A study of information seeking and retrieving. I. background and methodology. *Journal of the American Society for Information Science*, 39(3), 161-76.
- Saracevic, T. (2006). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II. In D. A. Nitecki, & E. Abels (Eds.), *Advances in librarianship* (vol. 30, pp. 3-71). London: Academic Press.
- Sarkar, D. (2009). Lattice: Lattice graphics (R package version 0.17-22) [Computer software]. Retrieved April 11, 2009, from <http://cran.r-project.org/web/packages/lattice/index.html>
- Savoy, J. (2005). Bibliographic database access using free-text and controlled vocabulary: An evaluation. *Information Processing and Management*, 41(4), 873-890.
- Sharp, E. C., Pelletier, L. G., & Levesque, C. (2006). The double-edged sword of rewards for participation in psychology experiments. *Canadian Journal of Behavioural Science*, 38(3), 269-277.
- Shiri, A., & Revie, C. (2005). Usability and user perceptions of a thesaurus-enhanced search interface. *Journal of Documentation*, 61(5), 640-656.
- Shiri, A., & Revie, C. (2006). Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. *Journal of the American Society for Information Science and Technology*, 57(4), 462-478.
- Sihvonen, A., & Vakkari, P. (2004). Subject knowledge improves interactive query expansion assisted by a thesaurus. *Journal of Documentation*, 60(6), 673-690.
- Sparck Jones, K., & van Rijsbergen, C. J. (1976). Information retrieval test collections. *Journal of Documentation*, 32(1), 59-75.
- Sparck Jones, K. (1974). Automatic indexing. *Journal of Documentation*, 30(4), 393-432.
- Sparck Jones, K. (1981). Retrieval system tests 1958-1978. In K. Sparck Jones (Ed.), *Information retrieval experiment* (pp. 213-255). London: Butterworths.
- Sparck Jones, K. (2000). Further reflections on TREC. *Information Processing and Management*, 36(1), 37-85.
- Sparck Jones, K. (2005). Epilogue: Metareflections on TREC. In E. M. Voorhees, & D.

- K. Harman (Eds.), *TREC: Experiment and evaluation in information retrieval* (pp. 421-448). Cambridge, MA: The MIT Press.
- Spink, A., Wolfram, D., Jansen, M. B. J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-234.
- Srinivasan, P. (1996). Query expansion and MEDLINE. *Information Processing and Management*, 32(4), 431-43.
- Sutcliffe, A. G., Ennis, M., & Watkinson, S. J. (2000). Empirical studies of end-user information searching. *Journal of the American Society for Information Science*, 51(13), 1211-31.
- Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science*, 37(5), 331-340.
- Svenonius, E. (2000). *The intellectual foundation of information organization*. Cambridge, MA: The MIT Press.
- Tague, J. M. (1981). The pragmatics of information retrieval. In K. Sparck Jones (Ed.), *Information retrieval experiment* (pp. 59-102). London: Butterworths.
- Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management*, 28(4), 467-90.
- Taylor, A. G., & Joudrey, D. N. (2008). *The organization of information* (3rd ed.). Westport, CT: Libraries Unlimited.
- TREC 2004 genomics track document set [Data file] (2005). Available from NIST TREC 2004 Genomics Track Web site, http://trec.nist.gov/data/t13_genomics.html
- Turpin, A., & Scholer, F. (2006). User performance versus precision measures for simple search tasks. *Proceedings of the ACM SIGIR Conference*, 29, 11-18.
- Turpin, A. H., & Hersh, W. (2001). Why batch and user evaluations do not give the same results. *Proceedings of the ACM SIGIR Conference*, 24, 225-231.
- Understanding metadata. (2004). Retrieved November 16, 2007, from <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>
- Vakkari, P. (2002). Subject knowledge, source of terms, and term selection in query expansion: An analytical study. *Lecture Notes in Computer Science*, 2291, 110-23.
- Vakkari, P., Pennanen, M., & Serola, S. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing & Management*, 39(3), 445-463.

- van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). London: Butterworths.
- Vickery, B. C. (1970). Document description and representation. *Annual Review of Information Science and Technology*, 6, 113-140.
- Voorbij, H. J. (1998). Title keywords and subject descriptors: A comparison of subject search entries of books in the humanities and social sciences. *Journal of Documentation*, 54(4), 466-476.
- Voorhees, E. M., & Harman, D. K. (Eds.) (2005). *TREC: Experiment and evaluation in information retrieval*. Cambridge, MA: The MIT Press.
- Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5), 697-716.
- Voorhees, E. M. (2005). The TREC robust retrieval track. *SIGIR Forum*, 39(1), 11-20.
- Voorhees, E. M., & Buckley, C. (2002). The effect of topic set size on retrieval experiment error. *Proceedings of the ACM SIGIR Conference*, 25, 316-323.
- Voorhees, E. M. (2008). On test collections for adaptive information retrieval. *Information Processing & Management*, 44(6), 1879-1885.
- Wacholder, N., Evans, D. K., & Klavans, J. L. (2001). Automatic identification and organization of index terms for interactive browsing. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 1, 126-134.
- Wacholder, N., & Liu, L. (2006). User preference: A measure of query-term quality. *Journal of the American Society for Information Science and Technology*, 57(12), 1566-1580.
- Wacholder, N., & Liu, L. (2008). Assessing term effectiveness in the interactive information access process. *Information Processing & Management*, 44(3), 1022-1031.
- Warnes, G. R. (2008a). Gdata: Various R programming tools for data manipulation (R package version 2.4.2) [Computer software]. Retrieved April 11, 2009, from <http://cran.r-project.org/web/packages/gdata/index.html>
- Warnes, G. R. (2008b). Gplots: Various R programming tools for plotting data (R package version 2.6.0) [Computer software]. Retrieved April 11, 2009, from <http://cran.r-project.org/web/packages/gplots/index.html>
- Weinberg, B. H. (Ed.) (1989). *Indexing: The state of our knowledge and the state of our ignorance*. Medford, NJ: Learned Information.
- Wildemuth, B. M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, 55(3),

246-258.

Witten, I. H., Moffat, A., & Bell, T. C. (1999). *Managing gigabytes: Compressing and indexing documents and images* (2nd ed.). San Francisco: Morgan Kaufmann.

Witten, I. H., & Bainbridge, D. (2007). A retrospective look at Greenstone: Lessons from the first decade. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 7, 147-156.

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis. *Proceedings of the ACM SIGIR Conference*, 19, 4-11.

Zhang, X., & Li, Y. (2008). Use of collaborative recommendations for web search: An exploratory user study. *Journal of Information Science*, 34(2), 145-161.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? *Proceedings of the ACM SIGIR Conference*, 21, 307-14.

Appendix A: Consent Form

Consent Form

Subject ID: _____

Subject Name: _____

Email-Address: _____

Experimenter Name: _____

Time/Date: _____

THE EFFECT OF DOMAIN KNOWLEDGE, SEARCH EXPERIENCE AND CONTROLLED INDEXING ON SEARCH EFFECTIVENESS

Research purpose: Thank you for volunteering to participate in our research study. Our goal is to advance our understanding of the effect of domain knowledge, search experience and controlled indexing on search effectiveness within the context of online database searching. The results of the study will guide us to develop better information retrieval system.

Procedure: The study will be conducted at your working place. You will get eight topics randomly selected from a list of candidate search topics. For each topic, you will be instructed to consult an online vocabulary look-up aid, called the MeSH Browser, or to come up with additional terms on your own. You will be asked to search an information retrieval system to find as many relevant documents as possible about that topic. You will have up to 10 minutes to do searches for each topic.

The entire session will last about two hours and fifteen minutes. During the first 20 minutes, you will fill out a brief questionnaire, and you will also be given a set of instructions and asked to do one practice topic. The actual experiment will take about one hour and 40 minutes. The last 15 minutes of the session will include an interview in which you are asked to assess your searching experience and provide suggestions of what might make it better.

When you use the experimental information retrieval system, the computer system will keep track of your on-screen actions. The session will be audio taped and the observer will take notes during the session. You will also be asked for demographic data such as age, gender, level, main area of study and native language.

Confidentiality: Your name will not be linked to any of the results. All data, including results, notes and tapes will be used only for research purposes. Any report on this study will not refer to you by name.

Risks or Discomforts: You may feel pressured or nervous due to the test-like nature of the topic. You may also feel self-conscious about being observed and audio taped. Please remember that the questions are only a vehicle to stimulate searching the systems.

Benefits: Your participation in this study will advance the cause of science and give you genuine research experience.

Duration: The entire session will last two hours and fifteen minutes.

Participation is voluntary: Your refusal to participate will involve no penalty. You may discontinue participation at any time without penalty or loss of benefits to which you may be entitled.

Compensation: At the completion of the session you will be paid by a gift card worth \$50.00, with incentives up to \$10.00, based on the average number of relevant documents from the top ten search results of reformulated searches. If you wish to stop before finishing, you will be paid a pro-rated amount (e.g., \$5 for 30 minutes of participation).

For more information: If you have any concerns or require any further information, please contact Ying-Hsang Liu (yhliu@scils.rutgers.edu), Nina Wacholder (nina@scils.rutgers.edu), or Michael Lesk (lesk@scils.rutgers.edu). *If you have any questions about your rights as a research subject, you may contact the Sponsored Programs Administrator at Rutgers University at:*

*Rutgers University Institutional Review Board for the Protection of Human Subjects
Office of Research and Sponsored Programs
3 Rutgers Plaza New Brunswick, NJ 08901-8559
Tel: 732-932-0150 ext. 2104
Email: humansubjects@orsp.rutgers.edu*

I, [print name] _____ agree to the conditions set forth above.

[Signed] _____ Date _____

Appendix B: Experimental Guidelines, Sample Document Records and Search Help

Experimental Guidelines

(Note: Lines in italics are experimenter directions, not to be spoken.)

Preparation

Materials for each subject:

1. *Experimental guidelines*
2. *Consent form (2 copies, one for subject, one for experimenter)*
3. *Demographic form*
4. *Digital voice recorder*
5. *Pen and paper*
6. *Laptop or Desktop Computer, with Internet connection*
7. *RU Express Card*
8. *Data record sheet for experimenter*
9. *Answer sheets (9 per subject, 1 of 9 are for training)*

Digital voice recorder set-up:

1. *Voice recorder is set to HQ and dictation mode (meeting mode is also acceptable)*
2. *Position the voice recorder so that both the experimenter's and the subject's voices can be heard*
3. *Check that the voice recorder is ready*
4. *After the experiment, the voice record should be transferred to computer*

Computer set-up:

1. *Open IE Web browser on the laptop/desktop*
2. *Make sure the right display setting of the computer: 1024*768*
3. *Set up the screen recording software*
4. *Write down the subject ID, the experimenter's name and time/date on the consent form; write down the subject ID on the demographic form; write down the subject ID on the chosen topic forms*
5. *Seat the subject so you can both see the computer screen*

Greenstone set-up:

1. *Link to <http://www.scils.rutgers.edu/irgs/gsdll/cgi-bin/library/> and choose System version B (without MeSH) first for training.*
2. *Click on the upper right corner PREFERENCES icon and set up SEARCH PREFERENCES. Select Form search ADVANCED Form type with 4 fields option and select Search history display*

- 5 search history records option.*
3. *Make sure that Case differences is set to ignore case differences and Word endings is set to whole word must match.*
 4. *Click on set preferences icon and then click on the search option on the top menu bar.*

MeSH Browser set up:

1. Link to <http://www.nlm.nih.gov/mesh/MBrowser.html>

Reminder: Write down the exact time the session starts (when the experimenter says "Good afternoon" or whatever). The entire introduction should be finished no more than 20 minutes from this time.

Reminder: Experimenter needs to take notes during the experiments.

Orientation

Good morning/afternoon/evening.

Thank you for participating in our study. The goal of this experiment is to observe people engaged in the process of searching an online database.

Please read this consent form, and if you agree, sign both copies. One is for you and one is for me. After you sign the consent form, we'll turn on the voice recorder.

----- allow subject to read and sign the consent forms

Thank you.

----- turn on the voice recorder

First we need you to fill out a questionnaire about your background.

----- hand the pre-experimental questionnaire to the subject

Remember all information is confidential; your name will not be associated with anything you do in this experiment.

----- allow subject to fill out demographic form

Practice Topic

Now I will explain the task. Several biologists are interested in a particular topic and they want information about that topic. Your task is to find documents related to the topic on their behalf.

Before you start, we will give you a practice topic to help familiarize you with the task. Do you have any questions?

----- *answer questions (if no, continue...)*

----- *Always use non-MeSH terms first in training.*

----- *experimenter shows subject the answer sheet and explains the items on the sheet*

Ok, here is a sheet that has the practice search topic.

First, please read through the search topic. Wait for a few seconds. The NEED field is a description of the kind of material the biologists are interested in. The CONTEXT provides background information to help you better understand what kind of information they want.

Concept Analysis

For each topic, you will start by looking for terms in the search topics. Look for terms that you can use in search interface. *Show concept analysis form. Reach over and underline terms.* In the example, this search topic has 3 or 4 words that you can use. Note that there are many different ways to do this. There is no single right answer.

Does that make sense to you?

Additional Terms

Now we would like to come up with additional synonyms or related terms to express the ideas. Write them down on the form. In the

example, high blood pressure is a synonym for hypertension, so we added that here. And relative risk is related to risk factors. *(That's fine if you don't come up with any.)*

Do you have any questions?

MeSH Browser

To make this task easier, sometimes you'll use the system, MeSH Browser, an online vocabulary look-up aid, to help you come up with additional synonyms or related terms. Just enter a term in the search box and click on the [Find Terms with ALL Fragments](#) button. Write down any of these terms that look useful on the form.

Do you have any questions?

System Features

Before you actually start searching, let me show you how the system works. This system provides basic search functions like the ones you use in popular search engines or Rutgers library databases. You can type in word or phrase in the search box. Here is a sample document record. *Go over with them.*

Use terms you have written down to look for documents. Feel free to use other terms as you go along. Take a look at search results until you feel satisfied with them.

Here is Search Help, just in case you need it. *Go over with them.*

Do you have any questions?

Task Details

In a minute you're going to perform a search to find useful documents.

You may take *up to* ten minutes for each topic – we'll let you know when ten minutes are up. The time spent doing concept analysis and coming up with additional terms is included in the ten minutes for each search topic.

Subject will do the training topic.

Experimenter makes notes about behavior and records time for the task.

When you're done with the search, we will show you 2 documents and ask you to rate the relevance of each. *Let them do it.*

Some of these search topics are hard. If you don't understand, just do the best you can!

To help you concentrate, there will be a prize at the very end of the experiment if you find useful documents. You have the opportunity to earn extra \$10, if your searches find good documents. We will calculate the accuracy of your results based on top ten documents of all topics.

Do you have any questions?

OK. You've finished the training. Do you have any questions about the task or any other aspect of the experiment? (*continue if "no"*)

There will be eight more topics like this; sometimes you'll get the MeSH terms to help you and sometimes you won't. But you'll only get the prize at the very end.

----- Set usage.txt file on Greenstone to be blank.

Formal Search Topics

----- Present subject with real topics and assigned experimental system, 4 topics per system, and each topic at most 10 minutes. Experimenter should take notes of subject's behaviors. Experimenter should record time for each topic.

----- Experimenter may give assistance if subject requests it, but avoid making suggestions that may bias the subject's behavior.

----- When each formal search topic is done, do the following data processing:

- *Open Terminal software on Mac and connect to SCILS server ssh
yhliu@scils.rutgers.edu*
- *Change to root directory cd / ; Change to/cgi-bin directory cd www/lrsgs/gsdll/cgi-bin*

- Run Perl script `perl run_lib6.pl s01 39 ../etc/usage.txt ../tmp/sub01/` (where **s01**, **sub01** are the subject ID, **39** is the topic number) It usually takes less than one minute to process the search log.
- Save screen capturing file when each search topic is done.

----- After finishing all search topics, run **TREC_Eval** program:

- Open another Terminal software on Mac and connect to SCILS server `ssh yhliu@scils.rutgers.edu`
- Change to root directory `cd /` ;Change to /cgi-bin directory `cd www/irgs/gsdl/trec_eval-8.0`
Type in `./trec_eval -a ../tmp/04.judge.txt ../tmp/sub01/s01_last1 > ../tmp/s01eval`

----- After TREC_Eval program processing is done, save screen capturing file.

----- Conduct the following **post-experiment interview**. Please use data record form to take notes.

1. How hard was this task?
2. What factors affected the level of difficulty?
3. Did the abstract give you ideas for additional search terms? If so, please explain.
4. Did you find MeSH terms useful for the searching task? If yes, please explain.
5. How could a system be designed to make this kind of searching easier?

Closing

Thank you very much for participating. Here is your Rutgers Knight express card worth \$25. According to our system you earn an extra X dollars. Be prepared to say something nice if they get \$0.

Would you like to receive a copy of our report on the results?

-----if yes, make sure that email address is on the consent form

Reminder: After the experiment, experimenter needs to:

Collect the laptops, voice recorder and data files;

Put the consent forms, demographic form, and answer sheet and data sheet into the folder; converting voice recording file.

Sample Document Record

Authored By:	Dollery C;
Paper Title:	[PMID-7473518] Hypertension trial results: consensus and conflicts.
Source:	J Hum Hypertens 1995 Jun;9(6):403-8.
Publication Date:	1995
Abstract:	In severe and accelerated hypertension the benefits of treatment are clearcut. In patients < 60 years of age with mild hypertension the main benefit is reduction of stroke by about 40%. The death rate from stroke is declining in many affluent countries for reasons which can only be partly explained by mass treatment of hypertension. In the MRC trial in patients < 60 years old it took 2500 patient/years of treatment to save one stroke. If the number of strokes is declining for other reasons the number of patient/years to save one stroke may be increasing. In older patients the absolute benefit is greater because they suffer more stroke events and because treatment also reduces coronary events.
MeSH Terms:	[MeSH terms] Antihypertensive Agents/adverse effects; Cerebrovascular Disorders/etiology/prevention & control; Clinical Trials; Human; Hypertension/complications/*drug therapy; Hypertension, Malignant/drug therapy; Randomized Controlled Trials;

Note. MeSH terms were only seen half of the time in search topics for each participant.

Search Help

1. Basic Search

Both Hypertension and stroke will appear in Title

Search and display results in order

Word or phrase	(fold, stem)	... in field
Hypertension	<input type="checkbox"/> <input type="checkbox"/>	Title
and <input type="text" value="stroke"/>	<input type="checkbox"/> <input type="checkbox"/>	Title
and <input type="text" value=""/>	<input type="checkbox"/> <input type="checkbox"/>	Abstract
and <input type="text" value=""/>	<input type="checkbox"/> <input type="checkbox"/>	MeSH Terms

Either Hypertension or stroke will appear in Title

Search and display results in order

Word or phrase	(fold, stem)	... in field
Hypertension	<input type="checkbox"/> <input type="checkbox"/>	Title
or <input type="text" value="stroke"/>	<input type="checkbox"/> <input type="checkbox"/>	Title
and <input type="text" value=""/>	<input type="checkbox"/> <input type="checkbox"/>	Abstract
and <input type="text" value=""/>	<input type="checkbox"/> <input type="checkbox"/>	MeSH Terms

Hypertensive rat as phrase will appear in Title

Search and display results in order

Word or phrase	(fold, stem)	... in field
"hypertensive rat"	<input type="checkbox"/> <input type="checkbox"/>	Title
and <input type="text" value=""/>	<input type="checkbox"/> <input type="checkbox"/>	Title
and <input type="text" value=""/>	<input type="checkbox"/> <input type="checkbox"/>	Abstract
and <input type="text" value=""/>	<input type="checkbox"/> <input type="checkbox"/>	MeSH Terms

Either Hypertension or hypertension will appear in Title

Search and display results in order

Word or phrase	(fold, stem)	... in field
hypertension	<input checked="" type="checkbox"/> <input type="checkbox"/>	Title
and <input type="text" value=""/>	<input type="checkbox"/> <input type="checkbox"/>	Title
and <input type="text" value=""/>	<input type="checkbox"/> <input type="checkbox"/>	Abstract
and <input type="text" value=""/>	<input type="checkbox"/> <input type="checkbox"/>	MeSH Terms

The query hypertension and its morphological variants, such as hypertensive and hypertensions, will appear in Title

Search and display results in order

Word or phrase	(fold)	stem	... in field
hypertension	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Title
and	<input type="checkbox"/>	<input type="checkbox"/>	Title
and	<input type="checkbox"/>	<input type="checkbox"/>	Abstract
and	<input type="checkbox"/>	<input type="checkbox"/>	MeSH Terms

2. Advanced Search¹

It is recommended to use advanced search in the query box, as shown below, to construct complex queries.

Search and display results in order

Word or phrase	(fold, stem)	... in field
	<input type="checkbox"/> <input type="checkbox"/>	Full Records
and	<input type="checkbox"/> <input type="checkbox"/>	Title
and	<input type="checkbox"/> <input type="checkbox"/>	Abstract
and	<input type="checkbox"/> <input type="checkbox"/>	MeSH Terms

Or enter a query directly:

query box

¹ This searching manual is adapted from Don, K. (n.d.). MGPP: A search engine for XML documents [Electronic Version]. Retrieved June, 16, 2006 from http://www.greenstone.org/docs/mgpp_user.pdf

Search Examples:

[hypertension#si]:TI	Search for hypertension, stemmed and casefolded, in TI (Title) field
[hypertension NEAR4 rat]:TX	Search for rat within 4 words either side of hypertension, in TX (Full Records) field
[hypertension]:TI & [stroke]:AB	Search for hypertension in TI (Title) field and stroke in AB (Abstract) field
[hypertension#si WITHIN3 rat#si]:TI	Search for rat, stemmed and casefolded, within 3 words following hypertension, stemmed and casefolded, in the TI (Title) field

Query Syntax

The query syntax allows the following:

Boolean operators:

& AND **|** OR **!** NOT, with **()** for precedence

Term modifiers:

#i #c #u #s —this is stemming and casefolding

#i case insensitive

#c case sensitive

#u unstemmed

#s stemmed

Proximity searching:

“...”—phrase searching, a form of very strict proximity matching where the query terms must be in the exact order specified by the phrase.

NEARx—this is used to specify the maximum distance apart (x words) two query terms must be for a document to match. NEAR by itself defaults to 20.

WITHINx—specifies that the second term must occur within x words after the first term. Similar to NEAR but the order is important. The default is 20.

Appendix C: Pre-Search Searcher Background Questionnaire

Pre-Search Questionnaire

BACKGROUND INFORMATION:

The information below is being collected for statistical purposes only. It will not be kept with any personally identifying information about you, and it will never be reported except in aggregate. Please complete the questions to the best of your ability. If you do not know an answer, or do not want to provide an answer, please leave the question blank.

1. What is your gender?

(please mark one)

☐ MALE

☐ FEMALE

2. What is your native language?

☐ English

☐ Other (please specify) _____

3. Have you ever taken any college-level biology class?

☐ No

☐ Yes

If yes, about how many?

_____ Undergraduate classes (please specify the number)

_____ Graduate classes (please specify the number)

4. Have you taken classes in how to do online searching? If so, about how many?

☐ No

☐ Yes, _____ classes (please specify the number)

5. Have you ever used Medical Subject Headings (MeSH)?

☐ No

☐ Yes

If yes, how much experience do you have?

☐ None ☐ A little ☐ Some ☐ A lot

6. Have you ever done scientific research for a class or on a professional basis?

☐ No ☐ Yes

If yes, describe briefly _____

7. Are you a student?

☐ No ☐ Yes

If yes, what is your current level of study?
(*please mark all that apply*)

☐ Freshman ☐ Sophomore

☐ Junior ☐ Senior

☐ Master (specify area of study) _____

☐ PhD (specify area of study) _____

8. Are you an information professional?

☐ No ☐ Yes

If yes, how long have you been working as an information professional?
(*please mark one box*)

☐ less than five years ☐ five to ten years

☐ ten to fifteen years ☐ more than fifteen years

9. For how many years have you used library online databases, such as MEDLINE or ERIC?

(please mark one box)

- | | |
|--|---|
| <input type="checkbox"/> not at all | <input type="checkbox"/> less than five years |
| <input type="checkbox"/> five to ten years | <input type="checkbox"/> ten to fifteen years |
| <input type="checkbox"/> more than fifteen years | |

10. How often do you search information using a search engine, such as Google or Yahoo?

(please mark one box)

- | | |
|--|--|
| <input type="checkbox"/> not at all | <input type="checkbox"/> several times a month |
| <input type="checkbox"/> several times a week | <input type="checkbox"/> every day |
| <input type="checkbox"/> several times a day or more | |

11. How old are you?

(please mark one box)

- ☐ younger than 18
- ☐ 18 or older and not yet 25
- ☐ 25 or older and not yet 35
- ☐ 35 or older and not yet 45
- ☐ 45 or older

Appendix D. Post-Search Search Perception and Comprehension Test for Each Search Topic

Post-Search Questionnaire

Search Topic ID: 39 Title: Hypertension Need: Identify genes as potential genetic risk factors candidates for causing hypertension. Context: A relevant document is one which discusses genes that could be considered as candidates to test in a randomized controlled trial which studies the genetic risk factors for stroke.
--

1. How difficult was the search task?
(please mark one number)

1	2	3	4	5
Not at all	Slightly	Fairly	Very	Extremely

2. How useful were the MeSH terms for the search task?
(please mark one number)

1	2	3	4	5
Not at all	Slightly	Fairly	Very	Extremely

3. Please indicate whether this document PMID-7988084 is relevant to the search topic.
(please mark one)

☐ Definitely Relevant ☐ Definitely Not Relevant ☐ I Don't Know

Authored By	Williams RR; Hunt SC; Hopkins PN; Wu LL; Lalouel JM;
Paper Title:	[PMID-7988084] Evidence for single gene contributions to hypertension and lipid disturbances: definition, genetics, and clinical significance.
Source:	Clin Genet 1994 Jul;46(1 Spec No):80-7.
Publication Date:	1994
Abstract:	Several large family studies are reviewed to identify results suggesting single gene traits contributing to the occurrence of hypertension in humans. Segregation analysis in families has suggested major gene effects for several highly heritable traits associated with hypertension. These

	include recessively segregating high sodium-lithium countertransport (major gene H2 = 34%), additively segregating low urinary kallikrein excretion (major gene H2 = 51%), and recessively segregating hyperinsulinemia (major gene H2 = 33%). In some families, hypertension and metabolic abnormalities (dyslipidemia, hyperinsulinemia, and obesity) seem to be related to several candidate genes studied but not conclusively proven (LPL deficiency mutations, dense LDL subfractions, or NIDDM with hyperinsulinemia). More recently, DNA markers have identified genes promoting hypertension. Glucocorticoid-remediable aldosteronism (GRA) promotes a rare but unusual form of hypertension that is unresponsive to ordinary medications but very responsive to glucocorticoid medications. GRA has been found in hypertensive persons with a specific mutation of the 11 beta-hydroxylase gene on chromosome 8q21. Many persons with essential hypertension carry a common 'susceptibility gene' at the angiotensinogen locus (chromosome 1q4) identified using linkage studies in siblings, association studies, and in studies of preeclampsia and hypertension in pregnant women. These first two well-established genetic loci promoting human hypertension represent two ends of a broad spectrum. The rare 'determinant' gene for GRA by itself seems to produce severe hypertension and early strokes. The angiotensinogen (AGT) 'susceptibility' gene is very common (30% of Utah Caucasians) and seems to predispose to hypertension but probably requires other genetic and environmental influences to be fully expressed.(ABSTRACT TRUNCATED AT 250 WORDS)
MeSH Terms:	[MeSH terms] Chromosome Mapping; Environmental Health; Human; Hyperlipidemia/diagnosis/*genetics/therapy; Hypertension/diagnosis/*genetics/therapy; Support, U.S. Gov't, P.H.S.; Syndrome;

4. Please indicate whether the document PMID-7802520 is relevant to the search topic.
(please mark one)

☐ Definitely Relevant ☐ Definitely Not Relevant ☐ I Don't Know

Authored By:	Hebert PR; Gaziano JM; Hennekens CH;
Paper Title:	[PMID-7802520] An overview of trials of cholesterol lowering and risk of stroke.
In:	Arch Intern Med 1995 Jan 9;155(1):50-5.
Publication Date:	1995
Abstract:	BACKGROUND: While blood cholesterol level predicts coronary heart disease, whether there is any association with the risk of stroke is unclear. Some, but not all, observational studies suggest that cholesterol level predicts risk of stroke, particularly ischemic stroke. This hypothesis is attractive because ischemic events constitute the vast majority of all strokes and, like coronary heart disease, involve atherogenic processes. METHODS: To investigate whether lipid lowering reduces the risk of stroke, we performed an overview of randomized trials that included more than 36,000 individuals. RESULTS: The mean reduction in cholesterol

	level in the treated as compared with the control subjects ranged from 6% to 23%. Those assigned to treatment experienced no significant reduction in all (fatal plus nonfatal) stroke (relative risk, 1.0; 95% confidence interval, 0.8 to 1.2) or fatal stroke (1.1; 0.8 to 1.6). CONCLUSIONS: The confidence interval for fatal stroke is wide, and alternative hypotheses, including either a small protective or harmful effect, cannot be excluded; however, the point estimates are compatible with no benefit of cholesterol lowering on the risk of stroke. Additional large-scale randomized trials assessing total mortality would more definitively address any benefits on stroke, as well as any excess nonvascular causes of mortality, for which risks of cholesterol lowering also remain uncertain.
MeSH Terms:	[MeSH terms] Brain Ischemia/complications; Cerebrovascular Disorders/etiology/*prevention & control; Cholesterol, Dietary/*administration & dosage; Human; Hypercholesterolemia/complications/*diet therapy; Randomized Controlled Trials; Risk; Treatment Outcome;

Appendix E. Post-Search Interview Questions

1. How hard was this task?
2. What factors affected the level of difficulty?
3. Did the abstract give you ideas for additional search terms? If so, please explain.
4. Did you find MeSH terms useful for the searching task? If yes, please explain.
5. How could a system be designed to make this kind of searching easier?

Appendix F. 20 Selected Search Topics

Topic #	Information Needs Statements
1	<p>Title: Ferroportin-1 in humans</p> <p>Need: Find articles about Ferroportin-1, an iron transporter, in humans.</p> <p>Context: Ferroportin1 (also known as SLC40A1; Ferroportin 1; FPN1; HFE4; IREG1; Iron regulated gene 1; Iron-regulated transporter 1; MTP1; SLC11A3; and Solute carrier family 11 (proton-coupled divalent metal ion transporters), member 3) may play a role in iron transport.</p>
2	<p>Title: Generating transgenic mice</p> <p>Need: Find protocols for generating transgenic mice.</p> <p>Context: Determine protocols to generate transgenic mice having a single copy of the gene of interest at a specific location.</p>
9	<p>Title: mutY</p> <p>Need: Find articles about the function of mutY in humans.</p> <p>Context: mutY is particularly challenging, because it is also known as hMYH. This is further complicated by the fact that myoglobin genes are also typically located in search results.</p>
12	<p>Title: Genes regulated by Smad4</p> <p>Need: Find articles describing genes that are regulated by the signal transducing molecule Smad4.</p> <p>Context: Project is to characterize Smad4 knockout mouse in skin (specifically skin) to establish signaling network. Identify all Smad4 targets to compare gene expression patterns of the knockout mouse to the normal mouse.</p>
15	<p>Title: ATPase and apoptosis</p> <p>Need: Find information on role of ATPases in apoptosis</p> <p>Context: The laboratory wants to know more about the role of ATPases in apoptosis.</p>
20	<p>Title: Substrate modification by ubiquitin</p> <p>Need: Which biological processes are regulated by having constituent proteins modified by covalent attachment to ubiquitin or ubiquitin-like proteins?</p> <p>Context: Ubiquitin and ubiquitin-like proteins have important roles in controlling cell division, signal transduction, embryonic development, endocytic trafficking, and the immune response.</p>
23	<p>Title: Saccharomyces cerevisiae proteins involved in ubiquitin system</p> <p>Need: Which Saccharomyces cerevisiae proteins are involved in the ubiquitin proteolytic pathway?</p> <p>Context: The researcher identified a protein in another yeast species and wants to compare it to the same one in Saccharomyces cerevisiae.</p>
27	<p>Title: Role of autophagy in apoptosis</p> <p>Need: Experiments establishing positive or negative interconnection between autophagy and apoptosis.</p> <p>Context: New information about experiments and genes involved in autophagic cell death.</p>

29	<p>Title: Phenotypes of gyrA mutations</p> <p>Need: Documents containing the sequences and phenotypes of E. coli gyrA mutations.</p> <p>Context: The laboratory has isolated some gyrA mutations in E. coli. They want to compare their mutant gyrA with the wild-type and other mutant sequences.</p>
30	<p>Title: Regulatory targets of the Nkx gene family members</p> <p>Need: Documents identifying genes regulated by Nkx gene family members.</p> <p>Context: The laboratory needs markers to follow Nkx family-member expression and activity.</p>
32	<p>Title: Xenograft animal models of tumorigenesis</p> <p>Need: Find reports that describe xenograft models of human cancers.</p> <p>Context: A xenograft animal model of cancer is one in which foreign tumor tissue is grafted into animals, usually rodents, providing a means to test various compounds for their ability to slow or halt tumor growth.</p>
33	<p>Title: Mice, mutant strains, and Histoplasmosis</p> <p>Need: Identify research on mutant mouse strains and factors which increase susceptibility to infection by Histoplasma capsulatum.</p> <p>Context: The ultimate goal of this initial research study, is to identify mouse genes that will influence the outcome of blood borne pathogen infections.</p>
36	<p>Title: RAB3A</p> <p>Need: Background information on RAB3A.</p> <p>Context: Further information about a gene is needed after it is identified through a gene expression profile. The genes are related to synaptic plasticity in learning and memory.</p>
38	<p>Title: Risk factors for stroke</p> <p>Need: Information concerning genetic loci that are associated with increased risk of stroke, such as apolipoprotein E4 or factor V mutations.</p> <p>Context: Candidate gene testing within a large Scottish case-control study of genetic risk factors for stroke. Future research includes investigations into other ethnically distinct populations.</p>
42	<p>Title: Genes altered by chromosome translocations</p> <p>Need: What genes show altered behavior due to chromosomal rearrangements?</p> <p>Context: Information is required on the disruption of functions from genomic DNA rearrangements.</p>
43	<p>Title: Sleeping Beauty</p> <p>Need: Studies of Sleeping Beauty transposons.</p> <p>Context: A relevant document is one that discusses studies on Sleeping Beauty. Interviewee's group studies a related element and want to know what others are doing in a similar field.</p>
45	<p>Title: Mental Health Wellness-1</p> <p>Need: What genetic loci, such as Mental Health Wellness 1 (MWH1) are implicated in mental health?</p> <p>Context: Want to identify genes involved in mental disorders.</p>
46	<p>Title: RSK2</p> <p>Need: What human biological processes is RSK2 known to be involved in?</p> <p>Context: After being identified via microarrays, the biological processes the</p>

	genes are involved in needs to be discovered.
49	<p>Title: Glyphosate tolerance gene sequence</p> <p>Need: Find reports and glyphosate tolerance gene sequences in the literature.</p> <p>Context: A DNA sequence isolated in the laboratory is often sequenced only partially, until enough sequence is generated to identify the gene. In these situations, the rest of the sequence is inferred from matching clones in the public domain. When there is difficulty in the laboratory manipulating the DNA segment using sequence-dependent methods, the laboratory isolate must be re-examined.</p>
50	<p>Title: Low temperature protein expression in E. coli</p> <p>Need: Find research on improving protein expressions at low temperature in Escherichia coli bacteria.</p> <p>Context: The researcher is not satisfied with the yield of expressing a protein in E. coli when grown at low temperature and is searching for a better solution. The researcher is willing to try a different organism and/or method.</p>

Source: TREC 2004 genomics track document set [Data file] (2005). Available from NIST TREC 2004 Genomics Track Web site, http://trec.nist.gov/data/t13_genomics.html

Curriculum Vita

Ying-Hsang Liu
yhliu@scils.rutgers.edu

Education

09/2002–05/2009	Rutgers, The State University of New Jersey PhD in Communication, Information and Library Studies
09/1997–06/2000	National Tsing Hua University Taiwan MA in Linguistics
09/1993–06/1997	National Taiwan University Taiwan BA in Library Science

Academic Appointments

Visiting Assistant Professor, Pratt Institute	08/2007–05/2008
Part-Time Lecturer, Rutgers University	09/2005–12/2005
Graduate Assistant, Rutgers University	10/2004–09/2007
Teaching Assistant, Rutgers University	09/2003–05/2005, 09/2008–05/2009

Professional Experience

Knowledge Express Technology, Taiwan	02/2001–02/2002
--------------------------------------	-----------------

Publications

- **Liu, Y.-H.** & Belkin, N. J. (2008). Types of query reformulation, search performance, and term suggestion devices in question-answering tasks. *Proceedings of the International Conference on Information Interaction in Context*, 2, 21-26.
- **Liu, Y.-H.** & Wacholder, N. (2008). Do human-developed index terms help users? An experimental study of MeSH terms in biomedical searching. *Proceedings of the American Society for Information Science & Technology Annual Meeting*, 45.
- Wacholder, N., Liu, L., & **Liu, Y.-H.** (2007). Selecting books: A performance-based study. *IEEE TCDL Bulletin*, 3(2). Available at <http://www.ieee-tcdl.org/Bulletin/v3n2/wacholder/wacholder.html>
- Wacholder, N., Liu, L., & **Liu, Y.-H.** (2006). User behavior during the book selection process. In *Proceedings of the American Society for Information Science & Technology Annual Meeting*, 43.
- **Liu, Y.-H.**, & Voon, W. (2005). An empirical-based taxonomy of collaborative technologies supporting intelligence analysis. In S. Hawamdeh (Ed.), *Proceedings of*

the 2005 International Conference on Knowledge Management (pp. 643-654). Singapore: World Scientific.

- Aakhus, M., Voon, W., & **Liu, Y.-H.** (2005) . Explicating tacit experiences in organizations: Evidence from online interns' discourse. In S. Hawamdeh (Ed.), *Proceedings of the 2005 International Conference on Knowledge Management* (pp. 41-52). Singapore: World Scientific.
- Wu, M.-M., & **Liu, Y.-H.** (2003). Intermediary's information seeking, inquiring minds and elicitation styles. *Journal of the American Society for Information Science & Technology*, 54(12), 1117-1133.