A Systematic Review of Interactive Information Retrieval Evaluation Studies, 1967–2006

Diane Kelly
School of Information and Library Science, 216 Lenoir Drive, CB # 3360, Manning Hall, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3360. E-mail: dianek@email.unc.edu

Cassidy R. Sugimoto
School of Library and Information Science, Indiana University, 1320 East 10th Street, LI 011, Bloomington, IN, 47405-3907. E-mail: sugimoto@indiana.edu

With the increasing number and diversity of search tools available, interest in the evaluation of search systems, particularly from a user perspective, has grown among researchers. More researchers are designing and evaluating interactive information retrieval (IIR) systems and beginning to innovate in evaluation methods. Maturation of a research specialty relies on the ability to replicate research, provide standards for measurement and analysis, and understand past endeavors. This article presents a historical overview of 40 years of IIR evaluation studies using the method of systematic review. A total of 2,791 journal and conference units were manually examined and 127 articles were selected for analysis in this study, based on predefined inclusion and exclusion criteria. These articles were systematically coded using features such as author, publication date, sources and references, and properties of the research method used in the articles, such as number of subjects, tasks, corpora, and measures. Results include data describing the growth of IIR studies over time, the most frequently occurring and cited authors and sources, and the most common types of corpora and measures used. An additional product of this research is a bibliography of IIR evaluation research that can be used by students, teachers, and those new to the area. To the authors' knowledge, this is the first historical, systematic characterization of the IIR evaluation literature, including the documentation of methods and measures used by researchers in this specialty.

Introduction

Within the last 15 years, online information retrieval has gone from a relatively specialized activity conducted primarily by professionals to an activity that, according to the Pew Internet and the American Life project[1], is conducted by nearly 87% of adult Internet users in the United States. Researchers have reacted to this development by focusing studies on users, their information needs and behaviors, and designing search user interfaces that support rich interactions. Research in this area, interactive information retrieval (IIR), blends research from information retrieval (IR), information behavior, and human computer interaction (HCI) to form a unique research specialty that is focused on enabling people to explore, resolve, and manage their information problems via interactions with information systems. IIR research comprises studies of people's information search behaviors, their use of interfaces and search features, and their interactions with systems. IIR research is also concerned with classification, indexing and retrieval techniques that are tailored to individual users or groups of users. One important characteristic of IIR research is that it focuses on people; it is common for users to be studied along with their interactions with systems. This is contrasted with classic IR research in which a user model might be used to guide research, but users are rarely studied directly. While IIR includes studies of information behavior, in this article we are primarily concerned with evaluation studies of IIR systems and interfaces.

Large portions of IR and IIR research are evaluations, in the form of experimentation or quasi-experimentation. Voorhees (2007) noted that the strong tradition of experimentation distinguishes IR research from other areas of computer science. In IR it is not enough to develop a new indexing or retrieval technique; one must also evaluate the technique. Classic IR evaluation has a prescribed and dominant method based on experimentation, which is rooted in the Cranfield studies (Cleverdon, 1960) and the Text

Received September 12, 2011; revised August 7, 2012; accepted August 8, 2012

 DOI: 10.1002/asi.22799

[1]http://www.pewinternet.org/Static-Pages/Trend-Data/Online-Activites-Total.aspx

REtrieval Conference (TREC; Voorhees & Harman, 2005). IIR studies, on the other hand, do not have prescribed experimental methods, but rely on a wide variety of methods and measures, perhaps because of the complexity of evaluating user behavior and system interfaces simultaneously. IIR researchers combine IR methods with methods from the behavioral sciences to create unique evaluation approaches.

The IIR evaluation method pioneered by participants in the Interactive Track at TREC (Dumais & Belkin, 2005) is perhaps the closest concept of a standard for IIR evaluations. This model was developed for a particular type of information search in a particular type of research setting (i.e., the laboratory). This model was also developed over 10 years ago. While it is still useful for some types of evaluations, there is growing dissatisfaction in the research community with the methods and measures that are available for, and used in, IIR investigations. In part, this dissatisfaction stems from insufficiencies of existing methods to evaluate the diversity of tools that are being developed. The TREC Interactive Track method, for example, is difficult to adapt to the evaluation of certain types of systems (e.g., adaptive) and certain types of search situations (e.g., exploratory) because it offers a temporally limited view of search behavior.

Another challenge in IIR evaluation is the development and use of measures. Although there have been a few measures developed specifically for IIR, in most cases measures comprise a mix of those developed to evaluate systems during batch-mode IR (e.g., recall, precision) and those developed to access general usability (e.g., efficiency, satisfaction). While many of the measures developed in TREC have undergone empirical and theoretical scrutiny, the underlying assumptions of these measures regarding the retrieval situation (i.e., noninteractive) do not always make them good fits to IIR. The usability measures are problematic because they are widely variable and usually created in an ad hoc and as-needed fashion by researchers. As a consequence, they are often of questionable validity and reliability.

As more researchers begin to focus on the design and development of systems that support IIR, there is an increased need for guidance about how to conduct evaluations and an increased need for valid and reliable measures that reflect a variety of interactive search situations. Maturation of a research specialty relies on its ability to provide standards for measurement and analysis, and on a firm understanding of past endeavors. Methods and approaches to evaluation are defining elements of research specialties (Valenza, 2009): domains are differentiated not only by the questions they can ask (cognitive coherence), but also the ways in which they can ask the questions (coherence of methods; Kuhn, 1996). The standardization of methods or "tasks" for a discipline can be used as an indicator of the type and status of a scientific field (Whitley, 2000). The presence and "coordination" of theory is another aspect that can be used to establish the maturation of a field of study (Whitley, 2000; Wagner & Berger, 1985). Lack of adequate theory has been presented as a shortcoming of other emergent information science fields (Vakkari, 2008). However, the degree to which theories, objectives, and methods in IIR have been established and, to some extent, canonized, is yet unstudied. This article presents a historical overview of 40 years of IIR evaluation studies using the method of systematic review. It will provide a foundation for describing evaluation in this domain. To our knowledge, this is the first attempt to systematically characterize this literature and document the methods and measures used in IIR evaluation research.

Background

Early Research in IIR

Although many people would classify IIR as a relatively new area because of the growth of online searching in the past two decades, Bourne and Hahn's (2003) historical account of online information services shows that users and interactions were prevalent themes in early information retrieval research. By the mid-1960s, Bourne and Hahn (2003) observed that several techniques had been introduced to assist users, including the display of an online thesaurus as search aid, the choice of novice or experienced searcher interface mode, the ability to save search queries to rerun at a later time or on a different database, relevance feedback, and systems prompting for further information about users' search requirements. Bourne and Hahn also describe the work of David Thompson, a researcher at Stanford University in the late 1960s. Thompson's work (described in his own work in 1971 with a review of interface design for IIR systems) focused on refining and testing theories about the human-machine interface and developing interactive man-machine dialogues.

Bourne and Hahn (2003) describe numerous research programs in academic, industrial, and governmental organizations that were focused on the development of interactive search systems, both operational and experimental, including DIALOG, ERIC, MEDLARS, LEXIS, and LEADERMART. Less attention is paid to experimental systems (except for commercial products). For example, Salton and colleagues' SMART system (Salton, 1971) is hardly discussed, although Bourne and Hahn (2003) acknowledge that this work contributed greatly to IR. Indeed, Salton is considered by most IR researchers as a pioneer who produced some of the most seminal and influential papers in IR (Spärck-Jones & Willett, 1997).

While the SMART research program is primarily referenced for contributions to indexing and retrieval, it also made contributions to IIR; the SMART system allowed users to successively broaden or refine searches and incorporated numerous relevance feedback techniques, which have arguably been the most popular interactive search techniques investigated in IR. Ide (1969), a member of the SMART team who investigated user interaction, remarked that "since the user's original query is often inadequate, some sort of user interaction with the retrieval operation is

DOI: 10.1002/asi

desirable" (p. 9). Ide proposed two types of interactive strategies: pre-search and post-search. The pre-search techniques assisted the user in constructing an initial query, while the post-search techniques operated on the search results and made use of relevance feedback. While Ide evaluated the effectiveness of the strategies, the evaluation approach was not interactive, which, of course, in the 1960s would have been difficult. However, this research acknowledged the necessity and importance of interaction along with the difficulties of evaluating IIR systems (Salton, 1970).

Many organizational activities occurred in 1971, which reflected the growing interest in interactive search systems. The American Society for Information Science chartered a new special interest group called User On-line Interaction (UOI) to encourage the development of online interaction models as they related to various user, computer, and information environments (Bourne & Hahn, 2003). A workshop entitled "The User Interface for Interactive Search of Bibliographic Data Bases," also held in 1971, was the first attempt to evaluate the accumulating research that was being conducted about interactive search systems (Walker, 1971). Participants of this workshop responded to a challenge paper issued by John Bennett (1971), in which he identified the key problems of interactive search, and focused the workshop on facilities for interactive search and complications caused by users with varying levels of computer expertise. Bennett further identified several challenges for the workshop including characteristics of the searchers, the conceptual framework presented to searchers, the role of feedback during search, and the role of evaluation and user feedback in the system redesign cycle.

Savage-Knepshield and Belkin (1999) used the design challenges identified by Bennett (1971) to review and analyze select IIR research that was conducted during three periods (mid-1960s to mid-1970s, mid-1970s to mid-1980s, and mid-1980s to late 1990s). With respect to IIR evaluation, Savage-Knepshield and Belkin found that both performance and user satisfaction were important in the early years; user satisfaction included coverage, recall, precision, response time, format, and effort. During the middle years, objective measures of recall and precision were dominant, and during the later years, subjective measures resurfaced and were closely related to standard usability measures.

Many researchers at the workshop considered their evaluations as case studies because they involved the study of a single system (usually operational). One concern was the extent to which findings from one case study generalized to another. Researchers at this time were keenly aware of the impact of situational variations in system use and effectiveness and the difficulties of identifying and accumulating findings that transcended a particular system. This concern was also expressed during a discussion session about user needs in which the discussion leader[2] began by stating, "I would like to start the discussion on user needs by stating my own position: user need studies are essentially useless" (Walker, 1971, p. 273). It is important to point out that "user studies" at this time were associated with surveys, most often conducted within a single institution, which sought to understand local users' information-seeking behaviors and needs. Siatri (1999) traced the first user studies in information science to works by Urquhart (1948) and Bernal (1948). These studies demonstrated an early interest by information professionals to develop services tailored to their users and an understanding of the difficulty of generalization.

Scope of IIR Studies

A variety of research areas and specialties contribute to IIR. In earlier work by Kelly (2009), the IIR research landscape was presented along a continuum, which identified and related different types of studies. This continuum will be used to situate the current work and narrow our focus. On one end of the continuum are system-focused studies that do not include test subjects, although a human might be involved in topic creation and result evaluation. The most common examples of these types of studies are those conducted within the context of the TREC and those conducted using resources created by TREC (Voorhees & Harman, 2005). On the other end of the continuum are studies that focus primarily on information-seeking behavior as it occurs naturally in a number of domains and across a variety of contexts (e.g., Case, 2002; Fisher & Julien, 2009). These studies do not necessarily focus on information seeking in electronic environments with search systems. Instead, these studies focus on a variety of sources and channels which a person might consult to resolve an information need.

Kelly (2009) described studies in the middle of the continuum as the classic or core IIR study. These are evaluation studies of IIR systems and interfaces that are conducted as experiments. Although no standardized methods and measures exist, many studies at this node occur in laboratories and incorporate elements of controlled experiments, coupled with questionnaires and other instruments from social sciences. In this article, we focus on the classic IIR evaluation study. Although many different types of studies contribute to the overall IIR research landscape, the evaluation studies form a core component where system-focused and human-focused approaches meet. It is important to recognize that not all studies of IIR are evaluations, although evaluation studies have historically been an important part of the IIR literature.

Two other types of studies that are discussed in Kelly (2009) are log analysis studies and experimental studies of information behavior. The distinction among these types of studies is especially important for demarcating the boundaries of the current research. While log analysis implies a method of data collection and analysis, such studies are uniquely situated on the continuum because of their importance and uniqueness. Although log analysis has been a central data collection technique for some time, industry researchers, especially, have introduced many new

[2]Davis B. McCarn from the National Library of Medicine, USA

DOI: 10.1002/asi

innovations in this area in recent years (Dumais, Jeffries, Russell, Tang, & Teevan, 2011; Jansen, 2009; Kohavi, Longbotham, Sommerfield, & Henne, 2009).

Researchers analyze search logs to discover behavioral patterns, and as an evaluation method using A/B tests and tests of interleaved search results (Dumais et al., 2011; Kohavi et al., 2009). *A/B test* is the term used to describe a live experiment when a slightly modified version of an interface, for instance, is distributed to a randomly selected number of users. The behaviors of these users are then compared to a set of users who function as a control group. Tests of interleaved search results are essentially live, within-subjects designs. For example, the results of two separate ranking techniques might be used to create a combined search results list through various interleaving techniques that is then presented to a user (Radlinski, Kurup, & Joachims, 2008). Researchers then examine clickthrough rates to see if results produced by one ranking technique garner more clicks than results produced by another. While A/B and interleaving tests can also be considered as examples of experimental studies of information behavior, a distinction can be made between large-scale log studies and smaller scale experiments of search behavior of the type that often occur in laboratories, use controlled tasks and systems, and gather other data in addition to log data (Kelly, 2009; Dumais et al., 2011). Furthermore, in these types of studies, people are aware that they are research subjects.

IIR Evaluation

Evaluation is one of the key features of IR and IIR research. Scriven (1991) defines evaluation as the systematic determination of the quality or value of something. In IR and IIR, this can be indexing, retrieval or ranking algorithms, user interfaces, or interactive techniques. Although an experiment is just one way a researcher might conduct an evaluation, this is the most popular and accepted method in IR and IIR.

Much has been written about IR evaluation and many research programs have been dedicated to the development of evaluation methods and measures. Harman (2011) and Sanderson (2010) provide the most current reviews of IR evaluation and measurement, and Robertson (2008) provides an informative retrospective. More historic viewpoints about IR evaluation can be found in the *Annual Review of Information Science & Technology* (ARIST). During ARIST's first six years, an article about evaluation was published every year. Of note, all articles discussed both system-centered and user-centered measures. King and Bryant (1971) published one of the first books about the evaluation of information services and products. It included chapters about evaluation of indexing techniques, the document screening process, user-system interface, and user surveys. The edited book by Spärck-Jones (1981) is considered a classic on IR evaluation; the chapters weave together both system-centered and user-centered issues. This book also contains one of the first guides to conducting IIR experiments with users (Tague, 1981).

Although interaction and users were a core part of the early discussions of IR evaluation, over time the user-centered and system-centered evaluation approaches diverged with some researchers focusing on evaluation of IR components (systems-centered), and others focusing on users and interaction (user-centered or human-centered; Cool & Belkin, 2011; Dervin & Nilan, 1986; Ingwersen & Järvelin, 2005; Järvelin, 2007; Saracevic, 1995). White and McCain (1998) showed by examining citation patterns that this split started in the 1970s and advanced more rapidly in the 1980s. This split coincided with the cognitive revolution, and more specifically with the Workshop on the Cognitive Viewpoint (DeMey, 1977), which is generally accepted as the official start of the human-centered perspective in IR (Ingwersen & Järvelin, 2005). Ingwersen and Järvelin (2005) state that although this viewpoint became associated with the user-centered approach, it really encompasses both user-centered and system-centered approaches because it considers "all information processing devices generated by man as well as information processes intended by man" (p. 25). This division into specialties also approximated the time that HCI began in earnest (Carroll, 2011); IR was a natural place to study HCI given the many operational and experimental IR systems in development and use.

One of the first compilations of IIR research can be found in Walker (1971), which includes over 150 citations to work relevant to IIR, including work on evaluation. Later, Belkin, and Vickery (1985) wrote one of the first books specifically about IIR, which included a lengthy discussion of IIR evaluation measures. Ingwersen and Järvelin (2005) discuss the evolution of IR and IIR evaluation and include an evaluation design framework and several examples of its use. Ruthven and Kelly's (2011) recent IIR book includes chapters describing methods for studying information behavior (Fidel, 2011) and evaluation (Järvelin, 2011). There are also a number of ARIST chapters that discuss aspects relevant to IIR evaluation, although many of these are outdated and focus more generally on information-seeking behavior or user-centered research.

There have been a few articles describing guidelines for conducting IIR evaluations (Borlund, 2003; Tague, 1981; Tague-Sutcliffe, 1992). Tague-Sutcliffe (1992) provides guidance about many aspects of evaluation including key measures and general study design. For many years, these articles were some of the only guidelines that existed about the IIR evaluation method, although Kelly (2009) published an updated IIR evaluation review. This article describes the design of IIR laboratory evaluations, including typical data collection instruments and measures. Borlund (2003) developed a framework for IIR evaluation which contained guidelines for more realistic evaluation, including, most notably, the simulated work task.

While user-centered research employs a large range of methods in comparison to system-centered research, less has been written on user-centered methods, especially in the

DOI: 10.1002/asi

context of IIR, and there have been relatively few studies devoted specifically to the development and evaluation of methods and measures. This is in contrast to IR (or, systems-centered research) where evaluation studies have occupied a portion of the literature. Furthermore, most of the published literature regarding IIR evaluation is outdated and, to the authors' knowledge, no one has used the method of systematic review to understand this research. The research reported in this article uses the method of systematic review to analyze 40 years of IIR evaluation studies to identify and characterize various aspects of the IIR evaluation method and, more generally, the body of research implementing IIR evaluations. Because systematic reviews are conducted infrequently in information science, in the next section, we provide a detailed explanation of this method and distinguish it from literature reviews and meta-analysis.

Systematic Reviews

Many readers will be familiar with literature reviews because they are a required component of most research reports. Cooper (1988) states that literature reviews comprise two main elements: "First, a literature review uses as its database reports of primary or original scholarship, and does not report new primary scholarship itself" and "Second, a literature review seeks to describe, summarize, evaluate, clarify, and/or integrate the content of the primary reports" (p. 107). Literature reviews provide information about what has been done in a domain and help establish that researchers have done the requisite background reading before proceeding with their own research. The importance of literature reviews is evident in the abundance of annuals and serials that exclusively publish literature reviews such as ARIST, *ACM Computing Surveys*, now's the *Foundations and Trends* series, and Morgan and Claypool's *Synthesis Lectures*. Cooper and Hedges (1994) state that literature reviews are among the most highly cited documents in the social sciences.

Some readers may also be familiar with specialized types of reviews such as systematic reviews and meta-analysis. Although researchers often use these terms interchangeably, it is important to distinguish among various types of reviews because each uses different methods. A literature review is the most basic type of review and involves the researcher gathering published reports that are relevant to a particular topic and summarizing, classifying, and qualitatively synthesizing them. The methods used to gather and analyze the literature are usually not described in detail. Readers have to trust that the researcher has used thorough and consistent methods to find and select articles, and that the researcher's coverage and treatment of each article is fair and balanced. Not surprisingly, literature reviews are often criticized because methods of construction do not adhere to the principles of the scientific method, and are instead at the discretion of the individual researcher (Wolf, 1986). Potential problems with traditional literature reviews include selective inclusion of studies (often based on the researcher's own impressionistic view of the quality of the study or limited searching abilities), lack of systematic analysis of studies, and lack of replicability (Wolf, 1986).

A systematic, or integrative, review attempts to make the review process more rigorous, explicit, and replicable (Cooper & Hedges, 1994). Researchers articulate a plan for gathering and analyzing studies and attempt to be exhaustive with their coverage of the literature. Researchers also take a neutral position during analysis and attempt to create generalizations from findings. Systematic reviews adhere to strict scientific guidelines to minimize potential selection and interpretation biases to ensure replicability (and hence reliability). The representativeness of the studies included in literature reviews, in particular, is a difficult issue because in most cases the parameters of the universe of studies are unknown (Feldman, 1971). Thus, one especially important characteristic of systematic reviews is that it is made explicit how studies are gathered, so that readers are in a better position to determine the representativeness of the studies and place boundaries on the conclusions made by researchers.

Systematic reviews comprise several major steps (Cooper, 1989):

1. State research questions.
2. Develop guidelines for collecting literature, including the specification of inclusion and exclusion criteria.
3. Develop a comprehensive search plan for finding literature.
4. Develop a codebook and code form for classifying and describing literature.
5. Code the literature.
6. Synthesize the literature.

Systematic reviews have a great deal in common with traditional content analysis, but are focused on the analysis of research literature.

According to Cooper and Hedges (1994), Glass (1976) coined the term meta-analysis and described it as "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p. 3). Meta-analysis is most often focused on outcomes rather than methods or theories and has attained prominence in medicine and health care (see, for example, the Cochrane Reviews[3]). Meta-analysis is considered a quantitative analysis technique because it typically comprises combining statistical results from a set of reports to determine what the cumulative evidence suggests. Many statistical techniques have been created for the sole purposes of combining statistics from multiple studies and comparing effect sizes. Another important aspect of meta-analysis is that quality standards are used to select studies for inclusion. Evidence can also be weighted to reflect various attributes of studies. Conversely, in systematic reviews, quality may or may not be an inclusion criterion depending on study

[3]http://www.cochrane.org

DOI: 10.1002/asi

purpose. If the goal is to present a snapshot of a body of research, then it is more important that both high-quality and low-quality studies are included because they represent the literature. Indeed, the inclusion of low-quality studies may highlight problematic practices.

Methods

A systematic review was used as the method for the present study, primarily because the focus was on methods and measures. It is also the case (as was learned later) that the literature of interest has very few reporting standards making meta-analysis infeasible. This is discussed in more detail later. This systematic review involved several steps:

1. Identify sources from which studies would be selected.
2. Develop and evaluate inclusion and exclusion criteria to guide the selection of articles from these sources.
3. Validate manual search and selection processes.
4. Develop a coding scheme for analyzing articles.
5. Apply coding scheme to articles.

These steps are described in more detail.

Step 1: Identification of Sources

The initial goal was to identify the sources from which studies would be selected. This meant identifying sources in which IIR studies were most likely to be published. Sources were first limited to journals and conference proceedings and needed to be peer-reviewed, such as scholarly publications, which eliminated trade magazines as well as some major conference proceedings, most notably, TREC and the initiative for the evaluation of XML retrieval (INEX). Although this meant the potential exclusion of some IIR evaluation papers, many of these papers were still included because they were later published in refereed publications. It is also the case that including these proceedings would likely lead to the overrepresentation of certain methods and measures, because research conducted as part of these evaluation exercises usually follow a standard, prescribed method.

It was determined that only full-length research papers would be included. Posters, short papers, demonstrations, book reviews and brief communications were excluded. In cases where a single study was published, first as a conference paper and then as a journal article, only the journal article was selected for inclusion. In most cases, overview papers that reported summary information about a series of studies were excluded because these papers usually lacked sufficient detail to allow for coding.

To identify sources, a list of journals and conferences that were believed the most likely to contain IIR studies was developed. The development of this list was informed by the inclusion and exclusion criteria for articles (described in the next section). For example, only studies that involved the evaluation of text-based searching were included, and so this meant that it was unnecessary to include multimedia conferences and journals.

Once the list of sources was compiled, four IIR experts were asked to review and comment on the list. These experts were Nicholas J. Belkin (Rutgers University, U.S.), Pia Borlund (Royal Danish School of Library and Information Science, Aalborg), Ian Ruthven (University of Strathclyde, U.K.), and Tefko Saracevic (Rutgers University, U.S.). Additional sources suggested by the experts were discussed and the list was modified accordingly. In total, 17 journals and 14 conference publications were included.

A 40-year time span from January 1967 to December 2006 was examined. These dates were chosen for a number of reasons. The first edition of ARIST was published in 1966, which indicates that the general area of inquiry was large enough to sustain a journal dedicated to reviews. The 1966 issue of ARIST contained chapters about evaluation (Bourne, 1966) and man-machine communications (Davis, 1966), and in each chapter some mention is made of studies that could be considered as IIR studies as defined here. These ARIST chapters indicate that many studies prior to 1967 involved time-shared computer systems and batch-mode submission requests by trained search intermediaries or were about filtering systems, none of which allowed for much interaction. It is also the case that many of the references in these two earlier ARIST chapters were to grey literature (e.g., technical reports), and it would have been difficult, if not impossible, to access this material.

In total, 2,667 journal units (where a unit equals an issue) and 241 conference units (where a unit equals a proceeding) were included in this study. Ninety-seven percent of the journal units ($n = 2592$) and 89% ($n = 206$) of the conference units were examined manually by the researchers. The failure to examine 100% of all units was because some units were unavailable or too costly to obtain. Most notably, several units of *Aslib Proceedings* are missing as are several older units of the Proceedings of the European Conference on Information Retrieval (ECIR) (despite a trip to the British Library). Table 1 details the dates of publication and number of units examined for each source.

Step 2: Development and Evaluation of Inclusion and Exclusion Criteria

The goal of this step was to develop and evaluate inclusion and exclusion criteria that could be used to systematically select articles for the study. The development of the inclusion and exclusion criteria began with a compilation of criteria that matched the target study type: IIR evaluations. This list was further refined as the authors independently applied the criteria to a 5% sample of units ($n = 147$). The physical or electronic units were examined manually and the title and abstract of each article were first reviewed. If a decision about inclusion could not be made based on this information, then the full text of the article was scanned. At periodic intervals the authors met to review the articles that had been selected and refine the inclusion and exclusion criteria as necessary.

DOI: 10.1002/asi

TABLE 1. Publications examined.

	Title	Earliest year examined	No. of units examined	Units not examined
Journals	*ACM Transactions on Computer-Human Interaction* (CHI)	1994	52	0
	ACM Transactions on Information Systems (TOIS)	1983	96	0
	Aslib Proceedings	1967	356	72
	The Canadian Journal of Information and Library Science	1976	75	0
	Information Processing & Management (IP&M)	1967	234	0
	Information Research: an international electronic journal	1995	47	0
	Interacting with Computers	1989	84	0
	International Journal on Digital Libraries	1997	23	0
	International Journal of Human-Computer Studies	1969	350	0
	Journal of Documentation	1967	177	0
	Journal of Information Retrieval (JIR)	1999	32	0
	Journal of Information Science (JIS)	1968	222	3
	Journal of the ACM (Association for Computing Machinery) (J. ACM)	1967	187	0
	Journal of the American Society for Information Science & Technology (JASIST)	1967	340	0
	Library & Information Science Research (LISR)	1979	112	0
	Online Information Review	1977	172	0
	New Review of Hypermedia and Multimedia	1989	33	0
Conferences	ACM Conference on Human Factors in Computers Systems (SIGCHI)	1981	26	0
	ACM Conference on Hypertext and Hypermedia (H&H)	1987	17	0
	ACM International Conference on Information and Knowledge Management (CIKM)	1993	14	0
	ACM International Conference on Intelligent User Interfaces (IUI)	1993	11	0
	ACM Special Interest Group on Information Retrieval Conference (SIGIR)	1971	30	0
	ACM/IEEE Conference on Digital Libraries (JCDL)	1994	18	0
	Annual Meeting of the American Society for Information Science & Technology (ASIST)	1967	31	8
	British Human Computer Interaction Group Annual Conference (BCS HCI)	1985	19	2
	Conceptions of Library and Information Science (CoLIS)	1991	4	1
	European Conference on Digital Libraries (ECDL)	1997	10	0
	European Conference on Information Retrieval (ECIR)	1989	12	16
	Information Interaction in Context (IIiX)	2006	1	0
	Information Seeking in Context (ISIC)	1996	4	2
	International World Wide Web Conference (WWW)	1994	9	6

Note. For journals, a unit represents an issue, and for conferences, a unit represents a proceeding.

The final list of inclusion and exclusion criteria is displayed in Table 2. Some of the most important criteria were that the study had to have human users engaged in interactive searching using traditional text retrieval systems in real-time. This excluded studies of filtering and human-mediated searching (some of the first IR studies involving users), and proactive information retrieval systems and recommender systems (because users do not always have to engage in searching). By searching, it is meant that the user has to do some querying (not just browsing) and engage in at least one exchange with the system. This criterion excluded many types of IIR studies including those where people only make relevance judgments of objects and those where users only formulate queries without viewing results.

Only studies of traditional text retrieval systems were included because studies of other types of information objects (e.g., audio, video) would make the study too large in scope. It is also the case that different types of information objects often require different kinds of evaluation methods that would complicate analysis and interpretation of the results. By "traditional" it is meant that searching occurs on traditional types of devices–mainly standard desktop or laptop computers. Studies excluded because of this criterion include those investigating search on mobile devices.

The study design was also considered in determining if a study should be included. The primary goal of the study had to be evaluation, two or more systems had to be studied, and at least one of these systems had to be an experimental system. It was acceptable if one system was a commercial search engine so long as there were two systems involved and one was experimental. The criterion that two systems needed to be studied excluded many usability studies of single systems. Large-scale log studies were also excluded unless the people were active participants in the research study (i.e., there was some intervention by, and interaction with, a researcher) and the other inclusion criteria were met (e.g., two or more systems were being compared). The studies also had to have some quantitative measures because one of the primary goals was to describe these measures.

Studies of information search behavior, which comprise an important part of the IIR literature, were excluded because they often involve more varied and nuanced methods and measures. For example, Joachims, Granka,

DOI: 10.1002/asi

TABLE 2. Inclusion and exclusion criteria for selection of study sample.

Class	Inclusion/exclusion criterion
Study purpose	• The purpose of the study should be to evaluate an IIR system or feature. This includes studies of various IIR retrieval techniques (where the system interface may be constant), studies of different interfaces and features, and studies of different interaction techniques. • The study should be empirical and attempt to use at least some aspects of the scientific method. • The study should compare two or more systems, where at least one of these systems is experimental (this criterion excludes usability studies of single systems and it also excludes studies where people are observed using commercial search engines only). • Subjects must engage in interactive searching, where it is necessary to enter a query and evaluate results. • If the study is declared a pilot by the authors, then it is excluded.
Study design	• Humans must be included as test subjects. Subjects should be adults with a normal range of cognitive abilities. • Only studies of end-users are included. Studies of mediated searching are excluded. • The study should report some quantitative results. These results do not necessarily have to be about the user (e.g., the results may be more focused on the system). • The study should take place in a controlled environment. Naturalistic studies are included if they meet other inclusion/exclusion criteria, but experiences must be bounded in some way and involve researcher intervention. • Studies only using log data where no intervention has occurred and users are basically unknown are excluded. However, studies are included in which the researcher introduces a variant of a commercial system, which pushes this to a subset of users; this use is compared with the use of the standard system. • Studies of evaluation design or measures are excluded, unless they are part of a larger IIR system evaluation. • Studies designed to collect only relevance assessments from users are excluded and studies of relevance assessment behavior are also excluded.
System type	• Studies of expert systems and decision support systems are excluded, unless such systems are for traditional IR applications. • Studies of filtering and recommender systems are excluded because such systems usually do not require users to engage in searching but instead push information to them. • Studies of online public access terminals are included. • Only studies using traditional desktop and/or laptop applications are included. Studies of mobile devices and other novel hardware are excluded.
Tasks and information objects	• Subjects must complete searching tasks when it is possible to enter a query and view results. • Search tasks can vary in kind and can be both natural and assigned. • Information objects searched must be textual documents. • Studies of multimedia retrieval systems are excluded. • Studies of e-mail management tools are excluded.

Pan, Hembrooke, and Gay (2005) conducted a study of search results selection behavior that involved varying the presentation order of search results. The goal of the study was not to evaluate a system, but rather to study users' selection behaviors. Another representative example is Vakkari's (2001) longitudinal study of students' search behavior when completing a research paper. This study did not focus on experimental search systems, but focused specifically on how subjects use existing systems and other information sources in completing a work task.

The criteria in Table 2 were applied manually to the 2,798 journal and conference units. A total of 127 papers were identified for inclusion in the data set. A paper functions as the coding unit for this study. While one might consider the 127 papers as the population of IIR evaluation studies, a more conservative approach is taken: It is likely that this number is close to the true number of IIR evaluation studies published during the time period, but it may be the case that several studies were excluded by either error or oversight. Thus, the 127 papers are considered as a representative sample that closely approximates the population of IIR evaluation studies. The list of papers included in the data set is available at http://ils.unc.edu/riire.

Step 3: Validation of Manual Search Process

The manual method of identifying studies was intensive and time-consuming. In many systematic reviews, keyword searches of literature databases are conducted to identify papers. This approach was considered infeasible for identifying IIR evaluation papers for a number of reasons. The first is that the validity of the searching relies on the creation of queries that will retrieve the majority of papers about the topic. It was not believed that such queries could be formed given the variety of labels that researchers attach to IIR evaluation papers. The second problem was that it was difficult to identify a set of representative databases that could be searched that would not result in the retrieval of a large number of false positives. However, to investigate the validity of the assumption that automatic search methods would be infeasible, results of the manual selection process were compared with the results of keyword searches using the Institute for Scientific Information (ISI; is now Thomson Reuters), Web of Knowledge (WoK)[4] and the Association for Computing Machinery (ACM) Digital Library[5].

[4]Searching conducted on May 21, 2009.

[5]Searching conducted on April 25, 2008.

DOI: 10.1002/asi

Two journals that had yielded a large number of results during manual searching, *Information Processing & Management* (IP&M) and the *Journal of the American Society for Information Science and Technology* (JASIST), were searched individually using ISI's *Science Citation Index Expanded*, *Social Sciences Citation Index*, and *Arts & Humanities Index*. Both were searched using previous and current titles (i.e., information processing and management *or* information storage and retrieval). The results were additionally restricted to "article" document types published between 1967 and 2006. The following query was entered for each journal individually[6]:

(users or user stud* or "interactive information retrieval" or user evaluation* or human subject* or interactive retrieval* or interactive IR or user interaction*)

The total number of documents retrieved from IP&M was 245. This list of 245 articles was then checked against the list of 17 articles identified through the intensive manual searching that were included in the study. Twelve of the 17 articles were found by ISI, meaning that 70.6% of the relevant articles would have been located by means of keyword searching and that 29.4% would have been missed had this been the only form of searching. The majority of the articles retrieved by ISI (95.1%) would have been false positives.

Using the same method for JASIST, 412 articles were identified. These were checked against the 34 articles from this journal chosen for inclusion in the study. The results were similar to that of IP&M: 67.6% of the documents would have been located by means of keyword searching and 32.2% would have been missed if this was the only form of searching. The majority of the articles (94.4%) would have been false positives.

A similar query was submitted to the ACM Digital Library (DL) targeting the Special Interest Group of Information Retrieval (SIGIR) Proceedings because the majority of studies selected manually were from this ACM publication. Unfortunately, the ACM DL did not provide information about how the search mechanism worked, so it was difficult to determine the effectiveness of the query. For instance, no mention was made about whether one could use truncation when querying or if stemming is used during retrieval. The results retrieved by the initial query did not make sense, so a modified version of the search query was created without truncation. Searching was also limited to the SIGIR Proceedings (excluding SIGIR Forum). The following query was issued:

((user or "user study" or "interactive information retrieval" or "user evaluation" or "human subject" or "interactive retrieval" or "interactive IR" or "user interaction") and (PublicationTitle:SIGIR)) and (not "SIGIR Forum"))

This query yielded 911 results. A query of the SIGIR Proceedings (using the last part of the query above) yielded 1,698 results. If one were to believe these results, then over half of the articles in the SIGIR Proceedings are related to the query that targets user studies and interactive searching. This result is puzzling because it seems much higher than it should be; not limiting the search to SIGIR Proceedings yielded 100,050 results, which further raises questions about how the search algorithm works, because in total the ACM DL contains 243,132 articles (according to the search feature).

The title and abstract of the first search result from the query that limited searching to the SIGIR Proceedings was examined to better understand how the search worked. This article was "An exploration of proximity measures in information retrieval," by Tao Tao and ChengXiang Zhai from SIGIR 2007. It is noted that none of the query terms appears in the title or abstract of this article, but the terms "information retrieval" and "use" were in the title and abstract respectively. Further inspection showed that the phrase operator was working, but that stemming was happening to some of the query terms, which completely changed the meaning of the term (e.g., user becomes use, which probably explains why the Tao and Zhai paper was retrieved). After viewing these results, it was also unclear which fields were being searched and there was no documentation describing this at the website. These findings, combined with an inability to limit the search to full papers and download the results list, demonstrated why this would not have been an effective approach to identifying IIR evaluation papers.

Step 4: Coding Scheme

For each paper in the data set, the features in Table 3 were coded manually. Features included article metadata such as

TABLE 3. Coding features used to analyze articles.

Class	Item
Publication	Year
	Source
Contributors	Name and affiliation
	Country
Study purpose	Research questions, hypotheses, and use of theory
	Objectives
Method	Subjects: number, type, label, compensation
	Corpus: type of documents, document source
	Search task: type, number, time
	Study design, type of data analysis
Measures	Output measures: conceptual and operational definitions
Cited works	Genres
	Years
	Source titles
	Item titles
	Cited authors

[6]The phrase *user study* was included in the query because many people refer to IIR studies as user studies. However, it is noted that the phrase user study originally described studies that investigated people's information-seeking needs (Siatri, 1999) and existed before there were interactive search systems.

DOI: 10.1002/asi

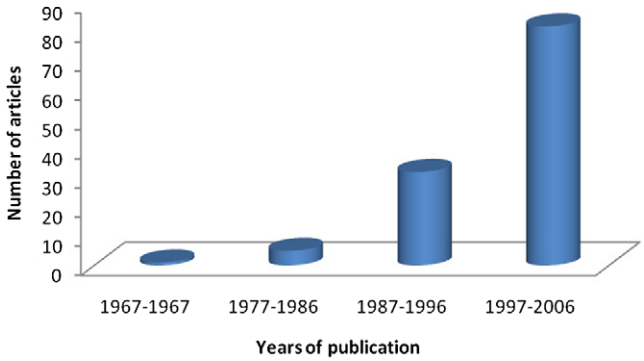

FIG. 1. Number of articles by years of publication. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

authors, affiliation, and publication year, purpose and objectives of the studies, use of theory, and methods used to conduct the evaluations, including tasks, instruments, and measures. Some of the features leant themselves naturally to systematic, quantitative coding. Other features required qualitative coding.

Results

One hundred twenty-seven articles were identified that matched the inclusion criteria. One hundred fifty studies were described in these papers and were coded (one article might report on more than one study). The Results section describes these articles/studies, organized by the feature scheme detailed in Table 3.

Characteristics of Publications

The majority of the articles (70%; $n = 89$) were published between 1997 and 2006. About 25% ($n = 32$) were published between 1986–1996; less than 4% ($n = 5$) were published between 1977–1986 and only one article identified for this study was published between 1967 and 1976. Figure 1 displays the distribution of articles across years of publication.

Fifty percent of the articles were found in three publications: JASIST, IP&M, and the SIGIR Proceedings (Figure 2), while the remaining 50% were scattered across 21 different publications. As listed in Table 1, 31 total publications were examined. Seven of the chosen publications had no IIR studies that qualified for inclusion in this study.

Contributors

A total 344 authors were associated with the articles. Of these, 261 were unique. On average, each article was authored by 2.71 people (median [*Mdn*]: 2; mode: 2; standard deviation [*SD*]: 1.50). The minimum number of authors was 1 and the maximum was 9. About 18% of the articles were written by a single author, 36% by two authors, 24% by three, and 22% were written by four to nine authors.

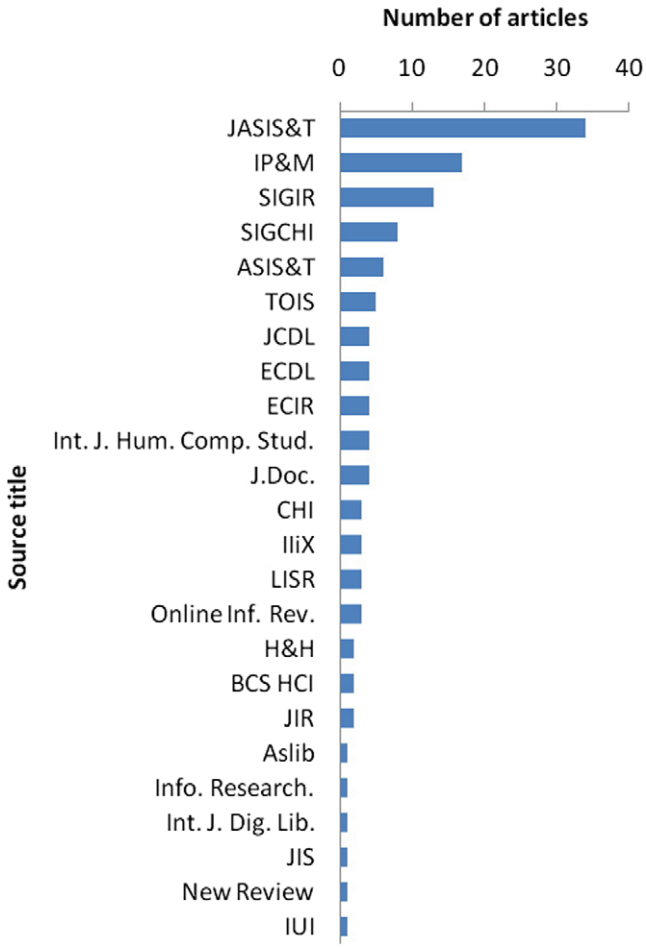

FIG. 2. Distribution of articles across publication sources. (See Table 1 for key to publication abbreviations.) [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Eighty-one percent of authors ($n = 212$) were associated with a single publication, 13% ($n = 33$) were associated with two publications, 3% ($n = 7$) were associated with three publications, 2% ($n = 4$) were associated with four publications, and less than 1% percent each were associated with five, six, and seven publications ($n = 2$, 2, and 1). Overall, about 6% of the authors contributed three or more publications to the data set.

Authors were associated with a total of 85 unique institutions: 70 academic institutions, eight corporate institutions, six government institutions, and one nonprofit institution. These articles represented 20 countries and the most frequently occurring countries were United States ($n = 77$) and the United Kingdom ($n = 23$). Figure 3 shows the distribution of various countries across the publications.

Study Purpose

Research questions, hypotheses, explicit use of theory, and objectives were coded for each of the documents in the corpus. Research questions were identified as objectives and purposes explicitly stated as questions, hypotheses were

DOI: 10.1002/asi

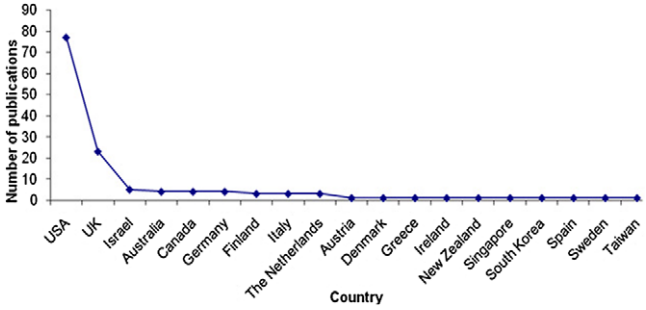

FIG. 3. Distribution of country affiliations. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

identified as those items explicitly labeled as hypotheses (such as, "it was hypothesized" or "the study was built on the following hypotheses"), and objectives were identified as both informal and formal statements of purpose (not stated as questions). To determine whether the research explicitly used theory, the front parts of the articles (i.e., the introduction and literature reviews) were examined for mention of specific theories that motivated the research. Articles were coded using the following categories: those that did not mention any theory, those that used theories and concepts from the information-seeking and behavior literature to motivate their research, those that used theories and concepts from other areas of study (e.g., psychology) to motivate their research, and those that explicitly used theory to generate hypotheses that would be tested in the research. One article might have several of these codes assigned to it.

Research Questions, Hypotheses, and Theory. Explicit research questions were found in 19.3% of the studies ($n = 29$), explicit hypothesis were found in 10.7% ($n = 16$) of the studies, and both a research question and a hypothesis were found in 4.7% of the studies ($n = 7$). In 65.3% ($n = 98$) of the studies, there was neither an explicitly stated research question nor hypothesis. For those studies with only a research question and no hypothesis, the average number of research questions for a given study was 2.5 (range of 1 to 5). In more than half of the studies with an explicit hypothesis, only one hypothesis was given. However, one study contained 16 hypotheses grouped into three categories. Those studies containing both explicit research questions and hypotheses contained, on average, 3.3 individual research questions. These studies typically had at least one hypothesis per research question, although one had as many as 12 hypotheses for seven research questions.

In total, there were 88 unique research questions found within the studies that contained research questions. These questions were grouped together by key terms contained within the questions. It should be noted that we are making no claim that the papers are *about* the key terms (i.e., effectiveness, usability, or relevance), but merely that these words (or derivatives/variations) were used within the question. The grouped terms and the instances of questions utilizing those terms are displayed in Table 4. Note that a research question could fall in multiple categories, if it utilized multiple key terms.

TABLE 4. Key words used in explicit research questions.

Term	No. of questions
Effectiveness	17
User preference and opinion	14
User behavior and interaction	12
Usability	11
Relevance	7
System performance	5
Time	5
Queries	4
User training and knowledge	3
Quality	3
Accuracy	1
Comparison	1
Reliability	1
Satisfaction	1
Theory	1
Unassigned	9

With respect to use of theory, most articles did not contain any explicit use of theory to motivate the research ($n = 81$, 64%). About 31% of the articles ($n = 40$) referenced at least one paper from the information-seeking behavior literature to motivate the research, while four articles ($n = 3\%$) contained reference to theories from other areas. Two articles (1%) mentioned both theories from information-seeking behavior literature and other fields.

Objectives. Although the majority of the studies did not contain explicit research questions or hypotheses, many implied or informal questions and hypothesis were embedded within purpose statements (such as "the primary goal of this" or "the objective of the experiment was"). An objective or purpose statement was identified for all of the documents. A keyword frequency was tabulated using Provalis Research's WordStat/QDA Miner to identify the most frequently occurring words within the purpose statements. The most frequent 15 terms are listed in Table 5.

Method

Subjects. Figure 4 displays the distribution of the number of subjects in each study. As a reminder, 150 studies were described in the 127 articles. In three cases, the authors did not report the number of subjects studied. The mean (M) number of subjects per study was 37.07 ([Mdn]: 24; mode: 24; [SD]: 40.81). The minimum number was 4 and the maximum was 283. Most studies that included large numbers of subjects were either conducted in corporate environments where the researchers were able to deploy the

DOI: 10.1002/asi

TABLE 5. Key words used in objective statements.

Word	Frequency
Search	95
System	76
User	58
Interface	41
Compare	34
Performance	33
Information	33
Effectiveness	31
Subject	29
Use	25
Retrieval	25
Evaluate	23
Query	21
Base	21
Investigate	19

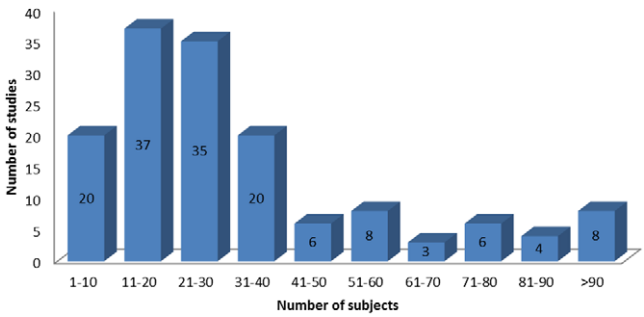

FIG. 4. Distribution of the number of subjects included in each study. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

experimental systems to many users or conducted with students enrolled in a university course.

Table 6 displays the types of subjects involved in the studies. This was difficult to code because many studies used mixed groups (e.g., computer science majors and communication majors; graduate and undergraduate students; employees of a corporation), and in many cases, little detail was provided. Therefore, the numbers presented in Table 6 do not sum to 150 because many studies reported more than one type of subject. Another important distinction is that subjects who are classified as library and information science (LIS) students or computer science students may or may not be majors–these groups include subjects who were enrolled in a subject-specific course.

The overwhelming majority of subjects were from academic settings and represented people who are generally well-educated. Few studies involved populations other than explicitly academic individuals; some examples include "women from suburban New Jersey," "People from the food industry," and "adult males from the Puget Sound region."

In more than 63% of the cases ($n = 95$), there was no indication of whether or not subjects were compensated. There were 12 reported cases in which participation was part of enrollment in a class, and seven cases in which employees were asked to participate as part of employment. In seven cases participation was voluntary and in one case a software gratuity was given as compensation. In 17 cases it was clear that there was some monetary compensation, but an exact amount was not given. There were only 15 cases in which an exact monetary amount was specified, ranging from $5 to $60. There were also cases in which a monetary reward was offered for fulfilling certain characteristics, for example "the two best performing participants across the treatments" or the "two best performing subjects based first on correctness and secondly on time" (ID74).

Corpus. The collection used in the study was described in 96% ($n = 144$) of the studies. Of these, TREC collections were used most frequently ($n = 39$; 27%). The web was also frequently used, with 15 (10%) studies using a "closed" or "artificial web" (static webpages that were crawled prior to the study or webpages with certain features, such as links, removed) and 20 (14%) studies using the "open web." Library collections (as accessed through the library catalog) were used in 7 studies (5%). Other test collections, such as Conference and Labs Evaluation Forum (CLEF) ($n = 3$) and INEX ($n = 1$) were also used. The remainder of the studies represented unique or individual collections.

In line with the type of collections used, the type of documents used included webpages or html documents ($n = 40$), news articles ($n = 25$), bibliographic records ($n = 17$), and (non-news) full-text documents ($n = 42$). These full-text documents included scholarly journal articles and text from monographs or textbooks. Many of the document types were implied, unspecified, or given only brief treatment in the text.

Search Tasks. Tasks were examined by whether they were assigned to the subject or of the subject's choice. Tasks were also examined for specified types or topics and the number of tasks each subject was asked to complete.

Type of tasks. Tasks were assigned (explicit tasks or topics given to subjects) in 134 studies and unassigned in four studies (e.g., instructions would read "browse for something of interest to you"). Most of the remaining studies were mixed—assigned topics were given, but the subjects were allowed to choose from among the specified topics. Task type was unspecified in 38 instances (25%) and largely varied in the remaining 75%. Ten studies identified the tasks as "simulated work tasks." Another 10 referred to the tasks as "search tasks," with an additional three as "retrieval tasks." Some studies differentiated tasks by levels of complexity (e.g., easy, medium, hard, difficult), and others by the process (e.g., browsing, searching) or by end product (known item search, citation task, essay task, fact search). In short, there was little standardization in implementing or reporting task types.

Number of tasks. There was wide variation in the number of tasks given to each subject largely due to the difference in

DOI: 10.1002/asi

TABLE 6. Characteristics of subjects studied.

Category	Group	N	Subgroup	N	Description	N
Academics ($n = 164$)	Students	144	Undergraduate	44	Unspecified	15
					CS	13
					LIS	8
					Psychology	3
					Business	2
					Engineering	2
					Chemistry	1
			Graduate	49	LIS	15
					Unspecified	14
					CS	9
					Medical	5
					Chemistry	2
					Psychology	2
					Economics	1
					Law	1
			Unspecified	51	Unspecified	27
					LIS	14
					CS	7
					Business	1
					Journalism	1
					Psychology	1
	Staff	7			Unspecified	5
					CS	1
					LIS	1
	Unspecified	5				
	Faculty	4			Unspecified	3
					CS	1
	Researchers	4				
Librarians ($n = 16$)	Unspecified	9				
	Academic	3				
	Staff	2				
	Professional	1				
	Special	1				
Employees ($n = 8$)	Corporate	6				
	Unspecified	2				
Other ($n = 27$)	Other specified populations	21				
	Unspecified populations	4				
	CS related populations	2				

Note. CS = computer science; LIS = library and information science.

task type (answering quiz questions or retrieving a citation could be seen as a less intensive search than writing an essay or completing a class assignment). The plurality of studies had between 6–10 tasks; and 76% of studies had 10 or less (Figure 5). The average number of tasks per study was nine ([*Mdn*]: 9; mode: 6). The largest number of tasks for any study was 56.

Time to complete tasks. Although time was used as a measure in many of the studies, only 38.7% ($n = 58$) of the studies reported an exact time limit given for each task within the method section. The average time was 17.1 minutes ([*SD*] = 13.19 min.) with a median and mode of 15 minutes. The shortest time limit given for any task was 1 minute and the maximum time limit reported was 75 minutes. These were primarily set as time *limits* only, with the possibility for the subjects to move more quickly if they desired; for example, many studies used wording similar to the following: "Participants were given up to 15 minutes to complete a task, but were allowed to end it when they felt they completed the tasks" (ID28). However, in one study, subjects were only "allowed to give up at any time after 15 min" (ID44). Some studies specified that there was

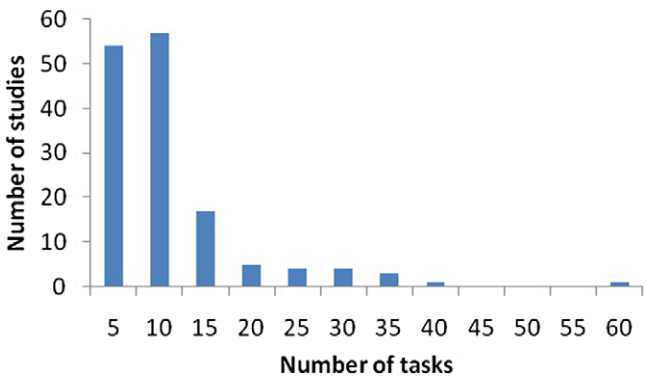

FIG. 5. Number of studies by number of tasks. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

DOI: 10.1002/asi

an "allotted time limit," but an exact number was not provided. Other studies reported a range of time or an aggregate time for all tasks. The time limit was reported as unlimited in 4% of the studies ($n = 6$). In 42% ($n = 63$) of the studies it was unspecified whether or not a time limit was given to the subjects.

There were many cases in which subjects were given instructions with implied times, such as to search "as quickly as possible" (ID226) or "until they felt they had exhausted their search strategies or had retrieved all relevant items" (ID59). In one study the number of relevant documents was given to the subjects, so they were allowed to move on as soon as they found all relevant documents or until the time limit had expired (ID224). In other studies subjects were informed that "a task would be considered fulfilled if 'enough' relevant items could be retrieved" (ID170). In some studies, with and without given time limits, subjects were provided with buttons such as "I don't know" (ID19) or "give up" (ID17) throughout the search process. Some subjects were told to search until they had tried a certain number of times and then to move on if they "had not yet retrieved anything relevant" (ID235). There were also cases in which subjects were given monetary incentives for finishing quickly (ID193).

Study Design and Method of Analysis

Given the criteria for gathering our data, the studies were all conducted within some type of lab or other controlled setting. Of the 150 studies, 149 gave an indication (either explicitly or implied) as to whether the study was a within-subject or between-subject study. One hundred and three (69%) were within-subject and 46 (31%) were between-subject studies. Typically one gets greater power with fewer subjects when using a within-subjects design, which might be why this type of design is used more frequently. It is also the case that this type of design allows for subjects to make cross-system comparisons and indicate preferences. Between-subject designs are more suited for situations where exposure to one condition will bias performance in another condition. These types of design are likely more prevalent in studies of search behavior, than system evaluation.

Most articles did not contain a specific section describing methods of analysis, so results were examined instead to determine which types of statistical tests were conducted. For studies where several different types of tests were used, each unique test was coded. For example, if an article reported results from an analysis of variance (ANOVA) and *t* test, then each of these tests were recorded. However, if an article reported results of multiple ANOVAs, then only one instance was recorded. The following numbers, then, are the frequency and percentage of articles containing a particular analysis method. In the majority of studies ($n = 57$, 45%), ANOVA was used as the method of analysis. This was followed by *t* test ($n = 33$, 26%), Mann-Whitey ($n = 11$, 9%), chi-square ($n = 8$, 6%), and Wilcoxon signed-rank test ($n = 6$, 5%). Correlation, Kruskal-Wallis and factor analysis were observed in fewer than 5% of the articles. Fifteen percent ($n = 19$) of the articles presented only descriptive statistics, while 9% ($n = 11$) did not provide any indication of which type of analysis was used, despite claiming statistically significant results or presenting probability values. Almost all the analyses were performed variable-by-variable and were conducted to compare the systems. Only a small percentage of articles described statistical analyses that attempted to model performance using multiple input variables ($n = 6$, 5%).

Measures

The coding and analysis of measures was more qualitative than expected because of the lack of clarity and consistency in the descriptions provided in the papers. An attempt was made to extract each measure at the lowest possible level and list the specific constructs that these measures were meant to represent. Unfortunately, most of the articles did not provide a system or framework for measurement and often did not indicate what particular measures (or signals) were indicated at the conceptual level. In many cases, signals were provided (such as number of queries issued) without much discussion of what they indicated and how their occurrence and frequency related to success or failure. Furthermore, many articles did not clearly specify the measures; the results sections of many articles had to be studied to identify measures.

In total, 1,533 measures were extracted from the 150 studies (again, note the difference between number of articles–127–and number of studies). Two studies did not describe any measures. Measures were classified according to conceptual class. The conceptual classes used were performance, process, and usability. These were derived inductively by studying the individual measures and considering the instances when researchers actually identified concepts.

Performance measures are measures that characterize how successful a subject is in accomplishing a specific task. Examples of measures include recall, precision, accuracy, correctness, and binary completion. Classic IR performance measures such as recall and precision are derived from another measure, relevance. While some researchers stated whether gold standard or individual relevance assessments were used, in many cases the methods used to determine relevance (and the instruments used by subjects) were not described in detail. It is assumed that binary relevance measures were primarily used (there were only a few examples of the use of multidimensional relevance).

Process measures are those that describe the interaction that takes place between the subject and the system. Examples include number of clicks, number of queries, and number of documents viewed. Time taken to complete the search task was classified as a process measure, although some researchers declared it as a performance measure. The reason for this is that its interpretation varied (as did most of the measures in this class). For some researchers, a greater

DOI: 10.1002/asi

TABLE 7. Descriptive statistics for measures.

Type of measure	*M*	*SD*	*Mdn*	Mode	Range
Performance	2.52	2.60	2	1	0–14
Process	3.44	4.68	1	2	0–33
Usability	4.34	5.85	2	0	0–28

Note. Statistics describing the number of different types of measures in each study; *M* = mean; *SD* = standard deviation; *Mdn* = median.

amount of time was positive, while for others it was interpreted as a negative. Most process measures were extracted from system logs. Process measures often provided the building blocks for efficiency measures, such as ratio of open to saved documents. Derived measures such as these were also categorized into this class.

Usability measures are those that characterize the subjects' perceptions of and attitudes about the system and their experiences using the system. Examples include usefulness of the system, user-friendliness, and satisfaction. These measures were primarily solicited from subjects using instruments such as questionnaires. When recording these measures, we listed every item identified by researchers; if researchers used a 12-item questionnaire, we recorded each item as a unique measure. There were a few cases where the researchers indicated that they used a specific usability questionnaire (the QUIS) or a multi-item questionnaire, but did not provide the individual items.

Descriptive statistics for the performance measures are reported in the aggregate in Table 7.

Performance Measures. Seventeen percent of studies reported no performance measures, while 25% reported one. As shown in Table 7, most studies reported a small number of performance measures. About 70% of the studies included between one and four performance measures.

The most frequently occurring performance measures were precision and recall, with some variations: for example, aspectual precision and recall, instance precision and recall, and interactive precision and recall. Aspectual precision and recall, as well as instance precision and recall, were measures used by the TREC Interactive Track, which most likely explains their frequent use in these studies. There were few occurrences of other classic IR measures such as mean reciprocal rank, f-measure, r-precision, and mean average precision.

There was a variety of precision measures used including precision at *n*, relative precision, retrieved precision, and viewed precision. These latter two measures, in particular, illustrate the difficulty in applying standard IR performance measures to interactive situations. Retrieved precision was defined as the total number of relevant documents saved divided by the total number of documents retrieved; retrieved precision is therefore a session-based measure, rather than a query-based measure (most classic IR measures are query-based, which makes them difficult to apply to interactive situations), and distinguishes between retrieving and saving. Viewed precision was defined as the total number of unique relevant documents saved divided by the total number of unique documents viewed. It accounts for the fact that subjects usually need to view documents before they can evaluate their relevance.

The computation of precision in system-centered evaluation is more straightforward because it is based on whether a document is retrieved or not; in user-centered evaluation, documents have to be retrieved, viewed, and marked relevant by a subject. This illustrates several problems using system-centered evaluation measures in IIR situations. Subjects often skip retrieved documents in the search results list that are considered relevant (at least according to baseline relevance assessments); including these unviewed documents in the computation of precision would not provide an accurate measure. Subjects also make decisions about which documents to view based on what they have already seen; documents that appear topically relevant might be skipped because they do not provide any new information. Because of these difficulties, researchers often reported two sets of precision measures: those based on subjects' relevance judgments, and those based on baseline relevance assessments. In the latter case, whether subjects viewed the document was immaterial.

There were fewer variations on recall. Most researchers used the standard recall measure of the proportion of relevant documents retrieved by subjects. Of course, there is still the problem of determining whether it is enough for a subject to retrieve a document or whether they need to view or even save the document for it to be included in the computation. Recall was reported less frequently than precision, which is most likely because one needs to know the number relevant in the collection for a given search task to compute this measure. One approach taken by several researchers was to create recall pools based on the documents retrieved by the subjects in their studies. This practice is somewhat limited because it depends upon the set of subjects and what is retrieved by a single system.

The computation of both precision and recall rely on relevance measures, which were often not described in the articles. In most cases, it was assumed that researchers' conceptualized relevance as topical, stable, and generalizable, operationalized it as a binary variable and used baseline relevance assessments. However, it was often unclear how subjects' behaviors (e.g., how they retrieved, viewed, saved, and printed documents) mapped onto this conceptualization. The most straightforward cases were where researchers simply used the number of documents saved by a subject as a relevance measure, and in many cases this measure was used by itself rather than as input for precision or recall. In a small number of cases, subjects were asked to explicitly evaluate the relevance of a document using a scale. Most often subjects were instructed to consider the topical relevance when making assessments, rather than pertinence or usefulness. In an even smaller number of studies, relevance was determined via the consensus method where a certain percentage of subjects needed to declare a document as relevant for it to count.

DOI: 10.1002/asi

A final issue related to the measurement of relevance was how to handle documents declared relevant by a subject, but not labeled as relevant in the baseline assessments. Although this is a common issue when using baseline relevance assessment in interactive evaluations, few papers reported the occurrence of such mismatches or discussed how such issues were resolved. Documents not marked relevant in the baseline assessments might not have been because they were not retrieved in the initial pool and were therefore never examined by the expert assessor, or they were examined by the expert assessor and declared not relevant.

In addition to the standard IR measures, another frequently occurring performance measure was correctness or accuracy. This measure was typically used when the search task had a specific answer, for example, fact-finding and known-item tasks. In some cases, researchers (or designated experts) used a scale to evaluate the quality of the answers subjects provided (for example, 0–5, where 0 *meant completely wrong* and 5 meant *perfect*). There were also a few instances where subjects were tasked with writing short essays using the information they found with a system. The essays were then graded by experts and used as a measure of system performance. Such assessments were rare and are based on work task performance rather than search task performance.

Process Measures. About 16% of the studies reported no process measures, while about 64% reported between one and four process measures. Similar to performance measures, most studies reported few process measures (Table 7). All of these measures were objective in the sense that they were not self-report measures.

The most frequently occurring process measures were time taken to complete a search task, number of queries issued, number of documents seen, and number of documents viewed. There was also a number of specialized counts based on the usage of features specific to a system. For example, number of suggested terms selected, number of documents marked as bad, number of times a help function was invoked. This class of measures included a large number describing queries including query length, number of AND operators, number of MESH headings used, number of misspelled terms, and number of unique query terms. This class also included a large number of time-based measures including time spent per query, time at which first relevant document found, dwell time, and time in state (querying, viewing documents, etc.).

As with the performance measures, frameworks for interpreting the measures were absent, which was even more problematic with this class of measures. For example, the interpretation of time taken, the most frequently reported process measure, is ambiguous and closely related to the task and user model (also often not reported). If a subject is given a fact-finding task and incidental learning is unimportant, then shorter completion times are likely better than longer completion times. However, if a subject is given an exploratory task and the purpose of the system is to facilitate discovery and incidental learning, then shorter completion times might indicate less learning or greater frustration with the system. The interpretation of time is further complicated by the fact that in many of these studies a fixed time limit was imposed. The interpretation of other process measures, such as number of queries entered and number of documents viewed, as well as how these measures relate to form a system of measurement, is also ambiguous and was often omitted from the articles. Without a measurement framework, it is difficult to understand what these measures tell us about the systems being evaluated, the quality of the interaction, and the experience of the user.

Usability Measures. The degree to which studies incorporated usability measures differed from that of performance and process measures; most notably, about 37% of the studies reported no usability measures. There was also greater variance in the number of the usability measures reported (Table 7). Most studies did not report any usability measures and about 30% reported between one and four measures, which seems like a low number considering the multidimensionality of usability and the general measurement principle that a number of items should be used to assess any one dimension when gathering self-report data to ensure reliability. The overwhelming majority (99%) of measures were subjective and comprised self-report data from subjects. There was also a small number of items that could be considered objective: number of errors made and time taken to learn a system.

Most usability items elicited absolute judgments from subjects using scales, although preference and other comparative judgments were elicited in a few cases. The most frequently occurring usability measures asked about satisfaction with the search results, ease of use, usefulness, understandability, ease of learning, general satisfaction, and time given to do the search. Many of these measures were further tailored to specific features of the system, for example: How satisfied were you with the layout of the interface? How useful were the term suggestions?

One of the major shortcomings was the extent to which these measures were described in the papers. Of all three types of measures, descriptions of usability measures were generally the poorest, and in many cases no descriptions were provided. This obviously has implications for reuse of these measures, but it is also problematic from the standpoint of assessing the quality of the results because these are tied directly to the quality of the questions asked and the choices and scales provided. Questions might be biased, loaded, or poorly worded; scale labels might be confusing or not represent a continuum. Developing self-report instruments requires a great deal of care, but much of the reporting practices we observed in these articles (including omission of the items and scales) suggested that these aspects were perhaps not taken seriously by many researchers.

Of all the measures, the interpretation of the usability measures was perhaps the most straightforward; generally, the more satisfied subjects are and the more usable they rate

DOI: 10.1002/asi

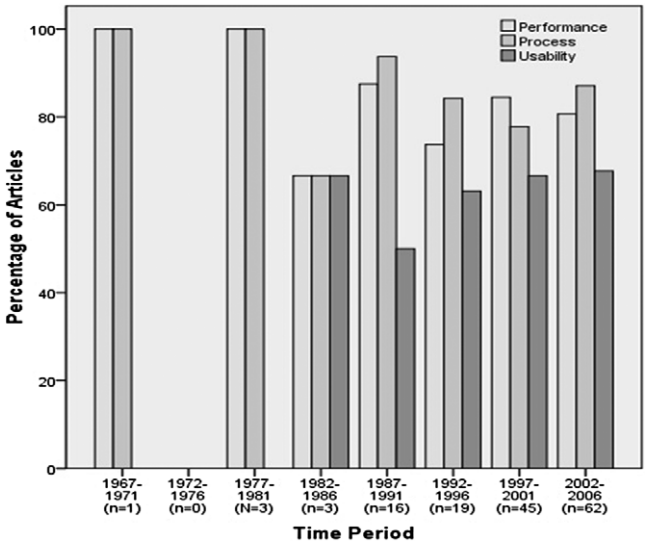

FIG. 6. Number of articles reporting each type of measures reported by time period.

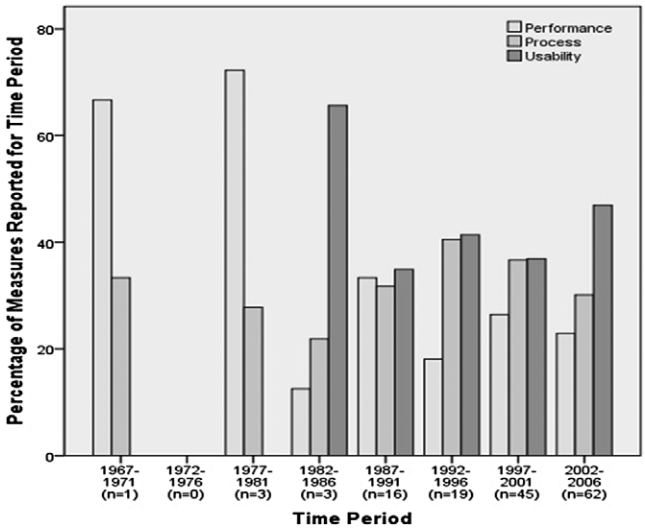

FIG. 7. Percentage of each measure reported by time period.

the system, the better. The biggest limitation of these measures was their development, which was mostly ad hoc and usually not described in detail. In most cases, the validity and reliability of the measures are questionable. There was also more variability in the items researchers used to assess usability than those used to assess performance and process, making it more difficult to do cross-study comparisons. Of course, some usability measures will be specific to the system being evaluated, but a common, core set of measures might allow the research community to provide a better trail that can be followed and used in the future.

Comparative Assessment of Measures Over Time. Of the 149 articles, 82% reported at least one performance measure, 85% reported at least one process measure, and 63% reported at least one usability measure. The percentage of articles from each time period reporting at least one performance, process, or usability measure is displayed in Figure 6. Because there were a very small number of studies observed during the early years, the number of studies is included to aid in interpretation. For example, in the first 5-year time period, 1967–1971, only a single article met the inclusion criteria for this study. This article reported at least one performance measure and one process measure. As the time progresses and the number of articles increases, the data provides more stable information. Usability measures first appeared during articles from the 1982–1986 time period, and remained a part of the measures through the last time period, 2002–2006. Although articles were more likely to include performance and process measures, not all articles reported these. In time period 1992–1996, only 74% of the articles reported performance measures.

Figure 7 displays the overall percentage of measures reported that belonged to each class of measure. The number of articles is again included to aid in interpretation, but the counts of total number of measures for each time period is based on the study rather than the article level. This figure shows that the proportion of usability measures to other measures has grown over time, while the proportion of performance measures has generally decreased. Although only three studies are included in the fourth time period, usability measures seem to have played an important part of these studies. It is, however, important to keep in mind that both the process and the usability measures were recorded at the item level.

TABLE 8. Types of cited material.

Type	Number	% of cited works
Serials	1,339	42.8%
Conferences	1,160	37.1%
Monographs	377	12.1%
Technical reports	81	2.6%
Web	75	2.4%
Theses	63	2%
Other	33	<1%

Cited Works

All references from each of the 127 articles were compiled into an Excel spreadsheet and coded by year, source type, source title, article/chapter title, and author name(s), and 3,128 references were identified, for an average of 24.6 references per article (min. = 3; max. = 77; [*Mdn*] = 22; mode = 22; [*SD*] = 13.96). The most dominant communicative genres of the cited works were journal articles (serials) (43%), conferences (37%), and monographs (12%) (Table 8).

Years

The years of the cited works were analyzed for all publication types (3,111 cited articles contained year

DOI: 10.1002/asi

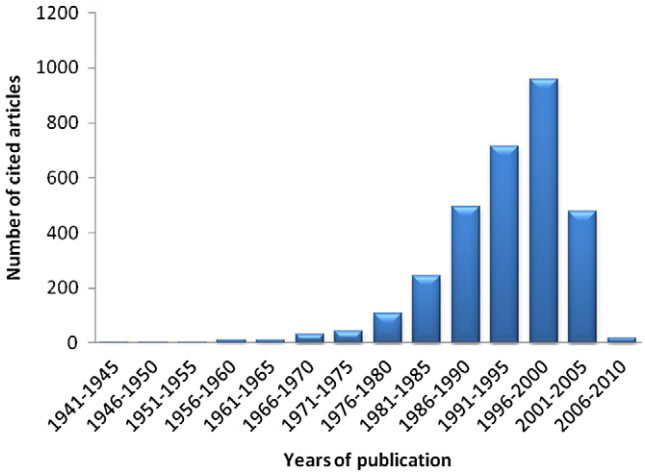

FIG. 8. Years of publication for cited articles. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 9. Age of the cited works, by genre.

	M	[*Mdn*]	Mode	Min.	Max.	*SD*
Conferences	5.46	4	3	0	42	4.70
Serials	8.08	6	4	0	60	7.50
Monographs	9.84	7	5	0	54	8.12

Note. *M* = mean; *SD* = standard deviation; *Mdn* = median.

information). As shown in Figure 8, the majority of citations were from articles published in the last 10 years. The earliest publication date was 1945[7] and the most recent publication date was 2006. The median year of publication was 1995 and the modal year was 1998. However, large differences can be seen by genre. Table 9 displays the age of the cited works, by genre.

Source Titles

In the analysis of sources used among the cited works, all webpages/websites were grouped together under the category "Web," and reports and technical reports were grouped under the heading "Reports." In addition, some sources comprised multiple journals or conferences that merged or changed name. A total of 730 unique sources was identified among the cited works. They were ranked in two ways: (a) the number of unique articles citing that source (in which each source receives a maximum of one count per article, regardless of how many times the source was cited within that particular article) and (b) the total number of citations across all sources (counting all times the source was cited, even within a single article). These analyses show both the degree to which a source is a staple reference across the entire body of literature and which sources may have been extremely popular with a small set of articles. The third analysis is used to demonstrate which sources have contributed large numbers of articles to the knowledge base, as opposed to a smaller number of highly cited articles.

The majority of sources were cited by only one article (68%; $n = 496$), were cited only once (65%; $n = 473$), and contributed one or fewer unique articles to the list of cited works (75%; $n = 549$). As shown in Table 11, IP&M, JASIST, and SIGIR were cited by a large proportion of the articles, with each cited by 60% to 66% of the article set. ACM Special Interest Group on Computer Human Interaction (SIGCHI) was also cited broadly, with citations from nearly half of the articles.

The 10 sources with the highest number of citations are shown in Table 11. The top four sources are the same as in Table 10, but the order within that group differs. In particular, IP&M has about two thirds of the total citations of either SIGIR or JASIST, despite being cited by the greatest number of articles.

Item Titles

A total of 1,891 unique articles, reports, chapters, or webpages/sites were identified among the cited works. A substantial majority (78%; $n = 1,474$) were found only once among the cited works. Table 12 shows those cited by at least 10 articles within our sample.

Table 13 shows the eight monographs that were cited by six or more articles. Totals include citations to the entire monograph and to sections within the volume.

Cited Authors

There were 2,488 unique cited authors in the corpus. These authors were ranked in two ways: (a) the number of unique articles citing that author (in which each author receives a maximum of one count per article, regardless of how many times the author was cited within that particular article) and (b) the total number of citations across all citations (counting all times the author was cited, even within a single article). The purpose of counting in both ways allows one to see both the degree to which an author is a staple reference across the entire body of literature and which authors may have been extremely popular with a small set of articles.

The majority of authors was only associated with a single article (67%; $n = 1,675$) and received only a single citation (62%; $n = 1,550$). Both rankings result in a long tail distribution, with many authors having relatively few citations/article associations and only a few having many citations/article associations. Those authors listed in the top 25 for both total number of articles citing that author and total number of citations is listed in Table 14 (the top 25 for both results in 28 unique authors).

Summary and Discussion of Findings

Sources, Authors, and Affiliations

Our systematic review indicates that this a relatively young field, with the majority of publications appearing in the

[7]There were three citations to Vannevar Bush's "As We May Think" article, which appeared in the *Atlantic Monthly* in 1945.

DOI: 10.1002/asi

TABLE 10. Most highly cited source titles.

Source	Type	No. of articles citing	% of all articles
IP&M; *Information Storage and Retrieval*; *Information Technology, Research and Development*	Journal	84	66%
JASIST; JASIS; *American Documentation*	Journal	81	64%
SIGIR	Conference	76	60%
SIGCHI	Conference	60	47%
Reports	Reports	54	43%
Communications of the ACM	Journal	42	33%
Journal of Documentation	Journal	36	28%
TREC	Conference	36	28%
International Journal of Human-Computer Studies; International Journal of Man-Machine Studies; Knowledge Acquisition	Journal	32	25%
Web	Webpages/sites	28	22%

Note. IP&M = *Information Processing & Management*; JASIST = *Journal of the American Society for Information Science and Technology*; JASIS = *Journal of the American Society for Information Science*; SIGIR = Special Interest Group of Information Retrieval; SIGCHI = ACM Special Interest Group on Computer Human Interaction; TREC = Text Retrieval Conference.

TABLE 11. Top sources by number of citations granted.

Source	Type	No. of citations	% of total
SIGIR	Conference	278	8.9%
JASIST; JASIS; *American Documentation*	Journal	273	8.7%
IP&M; *Information Storage and Retrieval*; *Information Technology, Research and Development*	Journal	185	5.9%
SIGCHI	Conference	156	5.0%
TREC	Conference	118	3.8%
Reports	Reports	89	2.8%
Web	Webpages/sites	80	2.6%
Communications of the ACM	Journal	76	2.4%
Journal of Documentation	Journal	66	2.1%
Annual Meeting of ASIST; ASIS	Conference	58	1.9%

Note. JASIST = *Journal of the American Society for Information Science and Technology*; JASIS = *Journal of the American Society for Information Science*; SIGIR = Special Interest Group of Information Retrieval; SIGCHI = ACM Special Interest Group on Computer Human Interaction; TREC = Text REtrieval Conference.

most recent time period studied. It is also a highly concentrated area of research, with 50% of all publications located in three publications: JASIST, IP&M, and SIGIR Proceedings. The work in this area appears to comprise small collaborative groups of two to three individuals per paper, with a small group of core researchers contributing multiple items. Authors are primarily from the United States and United Kingdom, although there was representation from a number of other countries. Academic institutions were the dominant affiliation of authors, with few authors from corporations and government institutions. The University of Strathclyde, Rutgers University, and University of Glasgow were leading institutions. Although academic institutions dominated the list of contributors, there were also contributions from corporations, such as Bellcore, IBM Research, Palo Alto Research Center, AT&T Labs, and Microsoft Research.

The skew towards academic research is likely the result of the type of study examined and time period; within the last 6 years, industry researchers have published a large number of studies about interactive search using log analysis and other methods (Dumais et al., 2011; Kohavi et al., 2009) and these studies were not included in this review. The skew is also likely a result of publication practices. Publishing is encouraged and rewarded in academia, but not necessarily in industry labs, and certainly some industry research cannot be published for proprietary reasons. Industry research is often published in technical reports and other forms of grey literature, which also might explain the low number of papers from industry and, more generally, the low number of papers from the earliest time period included in this study despite the extensive bibliography of pre-1971 IIR studies provided in Walker (1971).

Research Questions, Hypotheses, and Theory

The research was surprisingly void of explicit research questions and hypotheses—particularly given the experimental nature of this research. Less than 20% of the studies listed a research question and only 10% had hypotheses, and less than 5% had both. This raises serious implications as the primary IIR method is the experiment, where a research question and hypothesis are expected. Research questions were largely concerned with issues of the user (e.g., user preference, user opinion, user behavior, user interaction,

DOI: 10.1002/asi

TABLE 12. Most cited articles.

Title	Source	Year	Author(s)	No. of articles citing	% of articles citing
"Improving retrieval performance by relevance feedback"	JASIS	1990	Salton, G.; Buckley, C.	15	12%
"Finding facts vs. browsing knowledge in hypertext systems"	IEEE Computer	1988	Marchionini, G.; Shneiderman, B.	14	11%
"A case for interaction: a study of interactive information retrieval behavior and effectiveness"	SIGCHI	1996	Koenemann, J.; Belkin, N.J.	13	10%
"TileBars: Visualization of term distribution information in full text information access"	SIGCHI	1995	Hearst, M.A.	13	10%
"An evaluation of retrieval effectiveness for a full-text, document-retrieval system"	Communications of the ACM	1985	Blair, D.C.; Maron, M.E.	11	9%
"ASK for information retrieval: Part I. Background and theory"	*Journal of Documentation*	1982	Belkin, N.J.; Oddy, R.N.; Brooks, H.M.	11	9%
"Experimental components for the evaluation of interactive information retrieval systems"	*Journal of Documentation*	2000	Borlund, P.	10	8%
"Reexamining the cluster hypothesis: scatter/gather on retrieval results"	SIGIR	1996	Hearst, M.A.; Pedersen, J.O.	10	8%

TABLE 13. Most cited monographs.

Title	Year	Author(s)/editor(s)	No. of articles citing	% of all articles
"Information retrieval"	1979	van Rijsbergen, C.J.	14	11%
"Designing the user interface: Strategies for effective human-computer interaction"	1987, 1989, 1992, 1998	Shneiderman, B.	11	9%
"Introduction to modern information retrieval"	1983	Salton, G.; McGill, M.J.	11	9%
"Automatic text processing: the transformation, analysis and retrieval of information by computer"	1989	Salton, G.	8	6%
"Information retrieval: data structures and algorithms"	1992	Frakes, W.B.; Baeza-Yates, R.	8	6%
"Readings in information visualization: using vision to think	1999	Card, S.K.; Mackinlay, J.D.; Shneiderman, B.	7	6%
"Information retrieval interaction"	1992	Ingwersen, P.	6	5%
"Modern information retrieval"	1999	Baeza-Yates, R.; Ribeiro-Neto, B.	6	5%

usability) rather than the system. However, the objective statements revealed a slightly different aspect–the most highly ranked words are as follows (in order): search, system, user, interface, compare. This reflects the objective to "compare two search systems," which suggests an implicit research question focused on basic evaluation. However, the systems evaluated typically possessed experimental features that embodied some theoretical idea about relevance, retrieval, or interaction, or how it ought to occur. Formalizing these ideas as research questions would shift the focus from evaluation to interactive principles and possibly allow for more research depth in the field.

Few studies were motivated by formal theories, which is not too surprising because this review focused on evaluation studies. Instead, most studies were motivated by previous evaluations; researchers often used previous research to demonstrate the effectiveness or ineffectiveness of different techniques and to justify their own techniques. None of the studies explicitly started with a testable theory, developed hypotheses from the theory, and then used the results to modify or refute the theory. One likely reason for this is the lack of useful or testable theories to guide IIR evaluations.

Another reason is that these studies were focused on evaluation, which does not really require theories, except perhaps a theory of evaluation. A small number of studies made reference to ideas about or models of information-seeking behavior, but in most cases such reference was very general and functioned to demonstrate researchers' underlying beliefs about users. For example, in several papers the concept of anomalous states of knowledge (ASK; Belkin, Oddy, & Brooks, 1982) and the vocabulary problem (Furnas, Landauer, Gomez, & Dumais, 1987) were used to motivate system design. In other cases, researchers motivated their work using their own intuitions or ideas about users and behavior, although these were not presented as formal theories.

Very often theory is built from a body of empirical work that has examined a number of related hypotheses or from qualitative research. Because hypotheses were used in few studies and we did not examine articles that used only

DOI: 10.1002/asi

TABLE 14. Most cited authors.

Author name	No. of articles citing author (rank)	Difference between rank	Total # of citations (rank)
Salton, G	46 (1)	1	76 (2)
Belkin, NJ	42 (2)	1	90 (1)
Shneiderman, B	39 (3)	2	59 (5)
Saracevic, T	36 (4)	2	56 (6)
Croft, WB	35 (5)	1	56 (6)
Hearst, MA	34 (6)	4	51 (10)
Marchionini, G	34 (6)	3	62 (3)
Buckley, C	28 (8)	4	48 (12)
Harman, DK	28 (8)	13	34 (21)
Spink, A	28 (8)	0	53 (8)
Landauer, TK	27 (11)	3	43 (14)
van Rijsbergen, CJ	26 (12)	7	35 (19)
Dumais, ST	25 (13)	3	40 (16)
Card, SK	23 (14)	1	42 (15)
Jansen, BJ	23 (14)	6	53 (8)
Gomez, LM	22 (16)	5	34 (21)
Allan, J	21 (17)	0	39 (17)
Bates, MJ	21 (17)	11	30 (28)
Borgman, CL	21 (17)	4	34 (21)
Furnas, GW	21 (17)	4	34 (21)
Hancock-Beaulieu, MM	21 (17)	1	38 (18)
Robertson, SE	20 (22)	8	27 (30)
Chen, H	19 (23)	19	61 (4)
Lochbaum, CC	19 (23)	13	22 (36)
Ruthven, I	19 (23)	10	47 (13)
Hersh, W	18 (26)	15	49 (11)
Pirolli, P	16 (34)	15	35 (19)
Jose, JM	15 (38)	17	34 (25)

qualitative methods, this might explain why theories have yet to emerge from this particular subset of IIR research. In general, there are fewer studies that use qualitative methods in this specialty, increasing this number might provide more opportunity for theory development. Interestingly, IR theory was present in many of these articles, even if not explicitly presented or discussed as such, because it was captured and represented by the systems themselves through models such as language models, the vector space model, and latent semantic indexing.

While the distinction has been made between user-centered and system-centered evaluation approaches, we note that most of the user-centered evaluations we examined were actually very systems focused and followed a research model that closely resembled the traditional Cranfield/TREC model, only with users and some additional instruments and measures meant to track interactions and user experience. Most of the studies were primarily focused on demonstrating system effectiveness and few produced generalizable findings about search behavior or interaction. The exceptions were studies that examined the effects of individual differences on behavior and interaction, and low-level interface features such as the design of query input facilities or results snippets.

Finally, we observed some interesting parallels in our analysis to Vakkari (2008) who conducted a content analysis of papers published at the Information Seeking in Context Conference in 1996 and 2008. While the articles we examined do not belong to this specialty, research from this specialty could potentially inform system design because it is one of the main venues for information behavior research. This literature, in general, was dominated by empirical studies with few theoretical or methodological studies. Vakkari found an increase in the number of descriptive studies that used qualitative methods and found that the studies primarily contributed to the specialty by providing empirical support rather than new theories, concepts, or models (the contributions of about half the articles were characterized as "nothing special").

Vakkari noted that the applicability of this research to system design has decreased, perhaps explaining the disconnection we observed between the articles we examined and this literature. Vakkari also noted decreases in the number of explanatory studies and studies that had strong theoretical connections to past research, tested ideas, or generated models. Generally, he observed a narrowing of the research specialty and weakening of the ties to theory and system design. In our analysis, we noticed a convergence to a standard model of evaluation, a lack of connection to theory, and a focus on comparative assessment, rather than modeling and explanation. Some of this can certainly be contributed to the type of article we examined, but Vakkari's findings provide warning about the intellectual danger of research homogenization and a disconnect between theory and past research.

Methods: Subjects, Collections, and Search Tasks

Most studies had less than 30 subjects, with the plurality having between 11 and 20 subjects. The largest demographic of subjects were university students. In the majority of cases, it was unclear whether or not the subjects were compensated for their participation. The charge leveled against many academic studies can be seen here: Our studies demonstrate the needs of university students, but university students are not the only users of search systems. To truly create and evaluate systems with a diverse user base, the community must be cognizant of the populations used to assess these systems and work harder to include more diverse members of the community in laboratory studies.

No single collection was used in the majority of studies, although some standard collections (such as TREC, CLEF, and INEX) were reported. Many studies specified a "closed" or "artificial web" or individualized collections of news articles, bibliographic records, or other document types. This lack of standardization lessens the replicability of such studies, although it can be argued that these "natural" collections improve the external validity of the studies. Future work should be done to create standard data collections that reflect the adaptability and currency of systems. Most of the collections used were developed to support system-centered research, and future work might also create collections specifically designed for interactive searching.

DOI: 10.1002/asi

There was large variation in tasks, particularly in the reporting of task type. As with other features, this lack of standardization can be problematic, particularly in terms of replicability. Future research should review previous literature focused on classifying task type and label and justify tasks accordingly. Although task has always been an important component of IIR studies (because search tasks are necessary for people to exercise systems), in the last 10 years, increased research attention has been given to task, including how tasks should be designed for study (c.f. Li & Belkin, 2008). Task is now commonly used as an independent variable in many studies, but there is still much work to be done to develop task infrastructure for IIR evaluations or at least guidelines for creating tasks. Of course, the development of shared task sets or guidelines should be grounded in the understanding that the appropriateness of search tasks is closely related to the system and the tasks it has been designed to support.

Finally, it should be understood that search tasks are a part of the research method and therefore the tasks, and the development process for them, should be described in empirical reports.

Methods: Design and Analysis

Most studies were set up as within-subject studies, in which the subjects tested both the experimental and control system(s). Within-subject designs have many benefits including the ability to collect preference and comparative data from subjects and greater statistical power with fewer subjects. However, it can be difficult to rule out the possibility that any positive result supporting the experimental system is not an experimental artifact. Specifically, reactivity suggests that subjects might react more favorably to the system they presume the researcher is really studying. Subjects might also react more favorably to the experimental system because it is novel. While between-subject designs require more subjects, they allow the researcher to rule out many experimental confounds introduced by the method.

Because of the basic goals and design of these studies, the majority of researchers used either ANOVA or *t* tests to analyze the results. Despite using within-subjects designs, researchers often used tests that assumed independence (e.g., using an independent sample *t* test rather than a paired-sample *t* test), or at least did not specify otherwise. In some cases, the type of test conducted was not reported although statistically significant results were claimed or *p*-values were presented. There were a number of studies that relied only on descriptive statistics to make their claims, although these studies were primarily from the earlier time periods. Most of the analyses occurred on a variable-by-variable basis, although a few researchers used regression techniques to model performance using mostly individual difference variables such as search experience and occupation. None of the studies tried to model performance using behavioral variables such as queries entered and results examined. Instead these variables were used as dependent variables to demonstrate differences in the systems being studied. Although this is understandable because these studies were focused on evaluation, there is an opportunity in these data to generate more sophisticated models of search behavior.

Methods: Evaluation and Relevance Measures

This review found three basic types of evaluation measures: performance, interaction, and usability. Kelly (2009) grouped IIR measures into four classes: contextual, interaction, performance, and usability. In this study, we consider only output measures, so the contextual class was not applicable. Rather than interaction measures, we label a similar class of measures as process measures. Kelly (2009) distinguished between how usability has been defined and measured in IIR studies and how it is more traditionally defined and measured in the more general HCI literature. The ISO definition most often used in the general HCI literature considers usability to comprise effectiveness, efficiency, and satisfaction. The effectiveness measures include a mix of subjective and objective measures, including traditional IR performance measures such as recall and precision. The efficiency measures are similar to the process measures identified in this study and often come from system logs. Thus, while usability is the key construct in HCI research, in IIR research it is separated from performance and efficiency constructs, which is likely an artifact of IIR's close relationship to IR, in which performance and efficiency constructs are central. In IIR, the usability construct adds the piece that accounts for user attitudes and preferences, which in some ways limits thinking about the types of constructs that can be studied in these studies and fails to reflect the complexities of the user experience.

This review showed that there has been little innovation in the development of performance measures for IIR, including the development of more appropriate measures of relevance. The measurement of relevance has been a perennial problem in IR because the start of evaluation was at the heart of early debates about how performance should be measured (c.f. Cooper, 1973; Swanson, 1971). The lack of clarity of how this construct was defined and measured in the studies perhaps reflects researchers' desires to avoid this difficult issue. The measurement of relevance as well as any measure derived from it such as precision and recall necessarily relies on a clear explication of the concept. If one assumes that relevance is subjective and multidimensional, then the use of baseline assessment based on topicality is not particularly helpful. Moreover, it has been shown that IIR evaluation results can differ depending on how relevance judgments are measured (Vakkari & Sormunen, 2004), which questions the extent to which the findings from any of these studies can be compared.

System-centered performance measures deal with what Saracevic (2007) calls algorithmic and topical relevance. However, performance measures for IIR situations need to additionally deal with cognitive, situational, and motivational relevance. The application of system-centered

DOI: 10.1002/asi

performance measures to IIR situations in many ways has stymied research because they do not accurately reflect the user's experience. While researchers modified the classic measures in a variety of different ways, none of these efforts resulted in a new standard for IIR evaluation. Instead, the system-centered performance measures seem to have been accepted and maintained as part of the standard paradigm for valid IIR evaluations. There are new efforts and proposals, however, to create measures that capture the iterative nature of searching (e.g., Kanoulas, Hall, Clough, Carterette, & Sanderson, 2011) and that focus more on process and learning outcomes (e.g., Vakkari, 2010).

Citation Analysis

Articles in this area comprised more references over time, as is common with maturing research areas. Journals and conferences were largely relied on for referencing and most references were fairly recent (within the last 5 years). As self-citations are fairly common, it is not usual that IP&M, JASIST, JASIS, and *American Documentation*, and SIGIR would be among the most cited works. However, the study also demonstrated the importance of technical reports for communicating findings in this field, as 43% of articles cited this genre. The most cited articles show a lack of concentration in the works of influence on the field. No single article or monograph was cited by more than 12% of the articles. There does, however, appear to be a core set of authors influencing the field including (in rank order): Salton, Belkin, Shneiderman, Saracevic, and Croft. These authors are not necessarily the top contributors to this field, but their influence on the domain is undeniable. It is also interesting to note that two of the authors (Salton and Croft) are key contributors to IR research (not IIR research) and primarily conduct systems-centered research, which reinforces the interconnectedness of these two areas, and our earlier discussion about systems. Shneiderman is most noted for his contributions to HCI, which also demonstrates the interdisciplinary roots of IIR and the importance of human factors.

Reporting Practices

While conducting this study, we observed a large variety of reporting practices. One of our initial goals was to conduct a meta-analysis of IIR evaluation studies to investigate the relationship between performance measures and satisfaction. This was inspired by the works of Nielsen and Levy (1994) and Hornbæk and Law (2007), who conducted meta-analyses of the HCI literature to investigate the relationship between objective and subjective performance measures. Both studies found only medium to low correlations between these measures, and Nielsen and Levy (1994) further found that the majority of the systems studied were rated as better than average. In both studies, the researchers extracted descriptive and inferential statistics and were able to provide a quantitative synthesis of what had been found in the previous literature.

It was quickly realized that meta-analysis would not be possible in the IIR literature primarily because of varied, and sometimes poor, reporting practices: In many studies, researchers failed to report basic descriptive statistics such as means and standard deviations; in other cases researchers failed to report full results of inferential tests, including test statistic values, or values were represented only in bar charts or other figures, in which it is difficult to determine the precise value of the statistic. Given that researchers in IIR are from various academic backgrounds, it is somewhat unsurprising that no standard reporting practice has emerged. However, this study emphasizes the need to develop standard practices. Without such practices, it will be impossible to conduct meta-analysis in the future. No single study is definitive and the results of some studies conflict, while the same finding is replicated in others. Meta-analysis is a very useful method for systematically and reliably synthesizing research results across a body of literature, but, unfortunately, it is not possible to do such analyses at this time. Furthermore, standard reporting practices might make the literature more commensurate and amenable to the development of theories or theoretical constructs such as laws and models.

In our review of the research, we also found that many researchers published numerous papers using the same data set. As part of our inclusion and exclusion criteria, we selected only one paper to represent a project when we detected these situations. Such steps are necessary to avoid biasing the results to a particular research project by overrepresenting particular methods and measures. It was observed that researchers often did not acknowledge the related publications, or distinguish between secondary data analysis and piecemeal publishing. Moving forward, it would be useful for researchers to clearly distinguish between these two things and make explicit the relationship among a set of publications if they use the same data set. Journal publishers might consider requiring authors to reference and discuss past publications that used the same data set. This would make it easier for reviewers to evaluate the contributions of the new manuscript. Ultimately, such practices are important so that findings are not magnified simply because they are reported numerous times across different venues.

Study Limitations

This analysis was done with a very narrowly defined set of IIR papers and it is important to be mindful of this when drawing conclusions. It is not the case that the set of 127 articles represents everything ever written about IIR evaluation within the specified time period. Rather, this set only represents IIR evaluations that met the inclusion criteria (and even then errors and oversight are possible). It would be imprudent to make conclusions about the general state of IIR research based on the results of this study because many types of research were excluded (e.g., studies of search behavior and relevance and theoretical articles). However,

DOI: 10.1002/asi

evaluation studies comprise an important portion of the literature, so understanding more about the features of these studies can help researchers make more informed decisions in the future.

The period of study examined, 1967–2006, also has associated limitations. It is possible that some IIR evaluation studies were published before 1967. Because this study period ends in 2006, this paper cannot be considered a state-of-the-art review. Indeed, within the last few years, the use of online and remote experimentation has grown, new methods have been pioneered and proposed (especially by researchers at search engine companies; see Dumais et al., 2011; Kelly, Dumais, & Pedersen, 2009; Kohavi et al., 2009), and new measures are being developed (c.f. O'Brien & Toms, 2010).

A final limitation of this study is the potential reliability of the coding, especially that done with the measures. While coding and analyzing features such as the year of publication, number of authors, and number of subjects was relatively straightforward, the measures were not always obvious and it was often difficult to determine what the measures were and how they were computed. As noted in the results sections, there were many different indicators used, and it was not always clear what these indicators measured. This meant that the analysis of these measures was qualitative in nature and thus not as replicable as the analysis of some of the other features. With respect to validity, we were limited to what was explicitly reported by researchers in these articles. Researchers select which measures to report (sometimes favoring more interesting results) and are often constrained by page limitations. Thus, the measures identified in this study might not be completely representative of all the measures used in IIR evaluations.

Conclusions

This review has documented the evolution of the IIR evaluation method. Järvelin (2011) put it nicely when he observed that "information retrieval evaluation will not be remembered in history books for solving easy problems. Solving the difficult ones matters. Task-based and user-oriented evaluations offer such problems. Solving them can potentially lead to significant progress in the domain" (p. 137). While some standard practices have emerged over the past 40 years, there is still much work to be done on the development of methods and measures. These truly are the hard problems which, if addressed, will result in great benefits for the research community.

During this time, the research specialty of IIR has matured and has been reinvented. As more researchers begin to focus on the design and development of systems that support IIR, there is an increased need for guidance about how to conduct evaluations and an increased need for valid and reliable measures that reflect interactive search situations. With the maturation of the IIR specialty comes the ability to further specialize; it is hoped that research programs (whole or partial) will be dedicated to understanding the assumptions, limitations, and biases of existing methods and measures, and to the development and evaluation of methods and measures that are appropriate and useful to a range of IIR situations and contexts. As a first historical review of this literature, our study provides a baseline for the growth and maturation of the specialty.

One important outcome of this research is the IIR evaluation research bibliography. This bibliography of papers can be used as a reading list for graduate students or those new to the area, or as a starting point for an instructor compiling a reading list for IIR. Others can use this set of papers to explore, analyze, and reclassify the features that were coded in this study and to ask new questions about, and code new features of, this literature. Finally, this bibliography and the study results function as a snapshot of the literature that can be used as a benchmark and comparison point for future studies conducted in the upcoming decades. Now that the general characteristics of IIR evaluation studies have been documented, the community is in a better position to make decisions about how it wants to move forward in the areas of method, measures, and reporting practices. As Feldman (1971) observed, "A good integration, at the same time that it shows how much is known in an area, also shows how little is known. It sums up, but does not end. In this sense, it is only a beginning" (p. 100).

Acknowledgments

Thanks to Maxwell Felsher for his help preparing data files and creating the online bibliography.

References

Belkin, N.J., Oddy, R.N., & Brooks, H.M. (1982). Ask for information retrieval: Part 1. Journal of Documentation, 38(2), 61–71.

Belkin, N.J., & Vickery, A. (1985). Interaction in information systems: A review of research from document retrieval to knowledge-based systems. London, UK: The British Library.

Bennett, J.L. (1971). Interactive bibliographic search as a challenge to interface design. In D.E. Walker (Ed.), Interactive bibliographic search: The user/computer interface (pp. 1–16). Montvale, NJ: AFIPS Press.

Bernal, J.D. (1948). Preliminary analysis of pilot questionnaires on the use of scientific literature. In Proceedings of the Royal Society Scientific Information Conference (pp. 589–637).

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. Information Research, 8(3), 152.

Bourne, C.P. (1966). Evaluation of indexing systems. Annual Review of Information Science & Technology, 1, 171–190.

Bourne, C.P. & Hahn, T.B. (2003). A history of online information services, 1963–1976. Cambridge, MA: MIT Press.

Carroll, J.M. (2011). Human computer interaction (HCI). Retrieved from http://www.interaction-design.org/encyclopedia/human_computer_interaction_hci.html

Case, D.O. (2002). Looking for information: A survey of research on information seeking, needs and behavior. Lexington, KY: Academic Press.

Cleverdon, C.W. (1960). Report on the first stage of an investigation onto the comparative efficiency of indexing systems. Retrieved from http://www.sigir.org/museum/contents.html

Cool, C., & Belkin, N.J. (2011). Interactive information retrieval: History and background. In I. Ruthven & D. Kelly (Eds.), Interactive information seeking, behaviour and retrieval. London, UK: Facet Publishing.

DOI: 10.1002/asi

Cooper, H.M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. Knowledge in Society, 1(1), 104–126.
Cooper, H.M. (1989). Integrating research: A guide for literature reviews (2nd ed.). Sage Publications: Newbury Park, CA.
Cooper, H., & Hedges, L.V. (1994). The handbook of research synthesis. New York, NY: Russell Sage Foundation.
Cooper, W.S. (1973). On selecting a measure of retrieval effectiveness, part 1: The "subjective" philosophy of evaluation. Journal of the American Society for Information Science, 24, 87–100.
Davis, R.M. (1966). Man-machine communication. Annual Review of Information Science & Technology, 1, 221–254.
DeMey, M. (1977). The cognitive viewpoint: Its development and its scope. In M. De Mey et al. (Eds.), CC77: International Workshop on the Cognitive Viewpoint (pp. xvi–xxxi). Ghent, Belgium: University of Ghent Press.
Dervin, B., & Nilan, M. (1986). Information needs and uses. Annual Review of Information Science & Technology, 21, 3–33.
Dumais, S.T., & Belkin, N.J. (2005). The TREC Interactive Tracks: Putting the user into search. In E.M. Voorhees & D.K. Harman (Eds.), TREC: Experiment and evaluation in information retrieval (pp. 123–153). Cambridge, MA: MIT Press.
Dumais, S.T., Jeffries, R., Russell, D., Tang, D. & Teevan, J. (2011). Design of large-scale log analysis studies. Retrieved from http://research.microsoft.com/~sdumais/CHI2011-LogCourse-Share.pdf
Feldman, K.A. (1971). Using the work of others: Some observations on reviewing and integrating. Sociology of Education, 44(1), 86–102.
Fidel, R. (2011). Approaches to investigating information interaction and behaviour. In I. Ruthven & D. Kelly (Eds.), Interactive information seeking, behaviour and retrieval. London, UK: Facet Publishing.
Fisher, K., & Julien, H. (2009). Information behavior. Annual Review of Information Science & Technology, 43, 317–358.
Furnas, G.W., Landauer, T.K., Gomez, L.M., & Dumais, S.T. (1987). The vocabulary problem in human-system communication. Communications of the ACM, 30(11), 964–971.
Glass, G.V. (1976). Primary, secondary, and meta-analysis. Educational Researcher, 5, 3–8.
Harman, D.K. (2011). Information retrieval evaluation. San Rafael, CA: Morgan & Claypool Publishers.
Hornbæk, K., & Law, E.L.-C. (2007). Meta-analysis of correlations among usability measures. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (SIGCHI '07), San Jose, CA (pp. 617–626). New York, NY: Association for Computing Machinery.
Ide, E. (1969). Relevance feedback in an automatic document retrieval system. Retrieved from http://www.sigir.org/museum/contents.html
Ingwersen, P., & Järvelin, K. (2005). The turn: Integration of information seeking and retrieval in context. Dordrecht, The Netherlands: Springer.
Jansen, B.J. (2009). Understanding user-web interactions via web analytics. San Rafael, CA: Morgan & Claypool Publishers.
Järvelin, K. (2007). An analysis of two approaches in information retrieval: From frameworks to study designs. Journal of the American Society for Information Science & Technology, 58(7), 971–986.
Järvelin, K. (2011). Evaluation. In I. Ruthven & D. Kelly (Eds.), Interactive information seeking, behaviour and retrieval. London, UK: Facet Publishing.
Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In Proceedings of the 28th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '05), Salvador, Brazil (154–161).
Kanoulas, E., Hall, M., Clough, P., Carterette, B., & Sanderson, M. (2011). Overview of the TREC 2011 session track. In E. M. Voorhees & L. P. Buckland (Eds.), Proceedings of the 20th Text Retrieval Conference (TREC 2011). NIST Special Publication: SP 500-295.
Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval, 3(1–2), 1–224.
Kelly, D., Dumais, S., & Pedersen, J. (2009). Evaluation challenges and directions for information seeking support systems. IEEE Computer, 42(3), 60–66.
King, D.W., & Bryant, E.C. (1971). The evaluation of information services and products. Washington, DC: Information Resources Press.
Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R.M. (2009). Controlled experiments on the Web: Survey and practical guide. Data Mining and Knowledge Discovery, 18(1), 140–181.
Kuhn, T.S. (1996). The structure of scientific revolutions. Chicago, IL: University of Chicago Press. (Original work published 1962)
Li, Y., & Belkin, N.J. (2008). A faceted approach to conceptualizing tasks in information seeking. Information Processing & Management, 44(6), 1822–1837.
Nielsen, J., & Levy, J. (1994). Measuring usability: Preference vs. performance. Communications of the ACM, 37, 4, 66–75.
O'Brien, H.L., & Toms, E.G. (2010). Measuring interactive information retrieval: The case of the User Engagement Scale. In Proceedings of Information Interaction in Context (IIiX) (pp. 335–340). New Brunswick, NJ: ACM Digital Library. doi: 10.1145/1840784.1840835.
Radlinski, F., Kurup, M., & Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In Proceedings of the ACM Conference on Information and Knowledge Management (CIKM '08), Napa Valley, CA (pp. 43–52). New York, NY: Association for Computing Machinery.
Robertson, S.E. (2008). On the history of evaluation in IR. Journal of Information Science, 34(4), 439–456.
Ruthven, I., & Kelly, D. (2011). Interactive information seeking, behaviour and retrieval. London, UK: Facet Publishing.
Salton, G. (1970). Evaluation problems in interactive information retrieval. Information Storage & Retrieval, 6, 29–44.
Salton, G. (1971). The SMART retrieval system; experiments in automatic document processing. Englewood Cliffs, NJ: Prentice-Hall.
Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. Foundations and Trends in Information Retrieval, 4, 247–375.
Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development of Information Retrieval, Seattle, WA (pp. 138–146). New York, NY: Association for Computing Machinery.
Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. Journal of the American Society for Information Science and Technology, 58(3), 1915–1933.
Savage-Knepshield, P.A., & Belkin, N.J. (1999). Interaction in information retrieval: Trends over time. Journal of the American Society for Information Science and Technology, 50(12), 1067–1082.
Scriven, M. (1991). Evaluation thesaurus (4th ed.). Newbury Park, CA: Sage Publications.
Siatri, R. (1999). The evolution of user studies. Libri, 49, 132–141.
Spärck-Jones, K. (1981). Information retrieval experiment. London, UK: Butterworths & Co. Ltd.
Spärck-Jones, K., & Willett, P. (1997). Readings in information retrieval. San Francisco, CA: Morgan Kaufmann Publishers, Inc.
Swanson, D.R. (1971). Some unexplained aspects of the Cranfield test of indexing performance factors. Library Quarterly, 41(3), 223–228.
Tague, J.M. (1981). The pragmatics of information retrieval experimentation. In K. Spärck Jones (Ed.), Information retrieval experiment (pp. 59–104). London, UK: Butterworths & Co. Ltd.
Tague-Sutcliffe, J.M. (1992). The pragmatics of information retrieval experimentation, revised. Information Processing & Management, 28(4), 467–490.
Tao, T., & Zhai, C.-X. (2007). An exploration of proximity measures for information retrieval. In Proceedings of the 30th annual international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07), Amsterdam, The Netherlands (pp. 295–302). New York, NY: Association for Computing Machinery.
Thompson, D.A. (1971). Interface design for an interactive information retrieval system: A literature survey and a research system description. Journal of the American Society for Information Science, 22(6), 361–373.
Urquhart, D.J. (1948). The distribution and use of scientific and technical information. In Proceedings of the Royal Society Scientific Information Conference (pp. 408–419).

DOI: 10.1002/asi

Vakkari, P. (2001). Changes in search tactics and relevance judgments when preparing a research proposal: A summary of the findings of a longitudinal study. Information Retrieval, 4(3–4), 295–310.
Vakkari, P. (2008). Trends and approaches in information behavior research. Information Research, 13(4), 361.
Vakkari, P. (2010). Exploratory searching as conceptual exploration. In Proceedings of the Fourth Human Computer Information Retrieval Workshop, New Brunswick, NJ (pp. 24–27).
Vakkari, P., & Sormunen, E. (2004). The influence of relevance levels on the effectiveness of interactive information retrieval. Journal of the American Society for Information Science and Technology, 55(11), 963–969.
Valenza, R. (2009). Literature, language, and the rise of the intellectual disciplines in Britain, 1680–1820. Cambridge University Press.
Voorhees, E.M. (2007). TREC: Continuing information retrieval's tradition of experimentation. Communications of the ACM, 50(11), 51–54.
Voorhees, E.M., & Harman, D.K. (2005). TREC: Experiment and evaluation in information retrieval. Cambridge, MA: MIT Press.
Walker, D.E. (1971). Interactive bibliographic search: The user/computer interface. Montvale, NJ: AFIPS Press.
Wagner, D.G., & Berger, J. (1985). Do sociological theories grow. American Journal of Sociology, 90(4), 697–728.
White, H.D., & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. Journal of the American Society for Information Science, 49(4), 327–355.
Whitley, R. (2000). The intellectual and social organization of the sciences (2nd ed). New York, NY: Oxford University Press.
Wolf, F.M. (1986). Meta-analysis: Quantitative methods for research synthesis. Thousand Oaks, CA: Sage Publications.

DOI: 10.1002/asi