A Survey on Web based Interactive Information Retrieval

[1] Ratna Kumari Challa [2] B.Linga Murthy [3] K.Srinivasarao

[1] Assistant Professor, Dept.of CSE,JNTUKCE , JNTUK ,Kakinada
[2.] Software Engineer & Guide, IIIT - RKValley, RGUKT, Idupulapaya, Kadapa.
[3] Assistant Professor, Dept.of Computer Applications,YV University, ,Kadapa

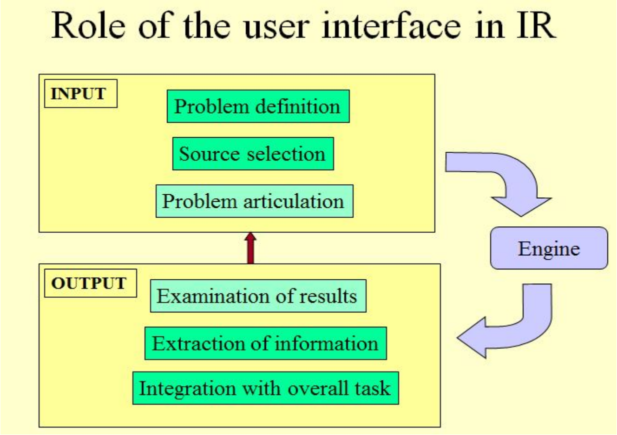
Abstract:Information retrieval is a fundamental component of human information behavior. The ability to extract useful information from large electronic resources not only is one of the main activities of individuals online but is an essential skill for most professional groups and a means of achieving competitive advantage. Information retrieval or search plays an important role in a wide range of information management and electronic commerce tasks. In spite of the importance of information retrieval, search systems are often poorly designed from a human computer interaction perspective. The goal of this paper is to articulate some of the opportunities and challenges in designing and evaluating highly interactive information retrieval systems called WEB.

I. INTRODUCTION

We are experiencing in our work and home environments a dramatic explosion of information sources that become available to an exponentially growing number of users. This has resulted in a shift in the profiles of users of online information systems: more users with no or minimal training in information retrieval (IR) have gained access to tools that were once the almost exclusive domain of librarians who served as intermediaries between end-users with their particular information needs and the information retrieval tools.

This situation has stimulated increasing interest in computerized tools that support end-users in their information seeking tasks. One important such situation is the information filtering (or *routing*) task, in which streams of information (such as email messages, newswire articles, or net news postings) are automatically filtered by a program based on specifications that are directly or indirectly obtained from the user.

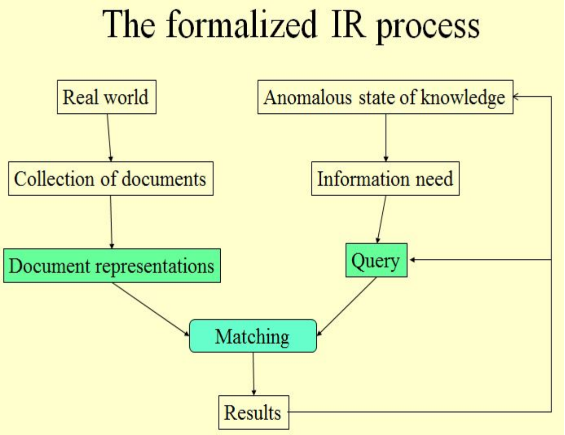
Users in all types of IR systems face the central difficulty of effective, interactive (re)formulation of queries which represent their information problems. Professional searchers using commercial IR systems have developed a variety of techniques and heuristics for addressing this difficulty in the context of Boolean query languages for exact-match, set-based retrieval from databases of indexed citations and abstracts of documents. Conversely, the difficulties faced by end-users with no training or experience in the use of these systems have been well documented. From experimental studies it has been known for some time that best-match; ranked-output retrieval techniques are in general superior in non-interactive settings to exact-match systems, such as commercial Boolean IR systems, in terms of recall and precision performance measures[2]. The following Fig1.1 shows the role of the user interface in IR.


Role of the user interface in IR

INPUT
Problem definition
Source selection
Problem articulation
Engine
OUTPUT
Examination of results
Extraction of information
Integration with overall task

Fig 1.1 Role of the user interface in IR

There have been many studies of user interaction with traditional Boolean systems (e.g. [5]) and some studies that have focused on novel interaction techniques (e.g. [1,7]). A few observational studies are concerned with relevance feedback [3, 4] but we are not aware of studies that have looked at relevance feedback in an experimental setting except for our own work in the context of the interactive track of TREC-3 [6]. The formalized IR process is shown in the Fig 1.2.


The formalized IR process

Real world
Anomalous state of knowledge
Collection of documents
Information need
Document representations
Query
Matching
Results

Fig 1.2 Formalized IR Process

A central question for the design of interactive systems in general is the amount of knowledge a user is required or expected to have about the functioning of the system and the level of control a user can exert. We share the "taskcentered” view that interfaces for the occasional user should hide as much as possible of the inner workings of a system and should instead present users with a view that focuses on the user's task. However, the question arises of *how much* knowledge and control a user should have in order to best interact with components such as relevance feedback that are central to the user task, here the formulation of an information need. At one extreme, the existence of such a tool can be completely hidden from the user: the set of “relevant” documents could be determined by some algorithm that takes as input a user's behavior such as the viewing, saving, or printing of documents. The other extreme would be a system that provides the user with complete control over the feedback mechanism: a user could provide lists of “good” documents to the mechanism, manipulate the query modifications (changed weights and added terms) suggested by the relevance feedback component, and even adjust internal parameters such as belief thresholds. Between these two extremes there is a large space of possible designs; the goal of this study was to explore this space through the design of four interfaces described in the next section.

II. A Traditional Iit: Interactive Information Retrieval Study

IIR studies are conducted so that we can better understand the human processesinvolved and conjointly, facilitate the design and development of improved searchsystems. IIR studies are distinct from analytical orsimulation IR approaches that result from manipulating system parameters but do notinvolve human participants interacting with a search system. IIR studies tend to vary in designfrom re-using existing data by analyzing existing log files (e.g., Catledge&Pitkow,1995; Jansen, Spink, Bateman &Saracevic, 1998), to observing human activities viatransaction logs (e.g., Hert&Marchionini, 1998), to observing humans in real-timesearch (Belkin et al, 2001; Hoelscher&Strube, 2000). We are primarily concerned with the last type: studies that explicitly examine humans actively participating in the searchprocess (e.g., Toms &Tague-Sutcliffe, 1996; Toms, 2000; Toms, Kopak, Bartlett &Freund, 2001; Toms, Freund & Yi, 2002). These may be studies in which the process isobserved (e.g., Bilal & Kirby, 2002) or studies in which one or more experimentalvariables are manipulated (e.g., Belkin et al, 2001; Toms, Kopak, Bartlett & Freund,2001).

A typical IIR experiment requires a number of steps that have been aptly described byTague-Sutcliffe (1992) in her excellent review of the design and conduct of an IIRexperiment:

a) Identify Purpose,
b) Define Variables,
c) Select Search Queries,
d) Select Database (i.e., SearchEngine/Interface),
e) Assign Treatments To Experimental Units,
f) Collect Data,
g) Analyze Data, And
h) Present Results.

Ten years later, some aspectshave changed. Search interfaces are being designed independently of the search engine and a mix and match approach of interface and database may be applied. Thesearch query is but one aspect of a search task, “the manifestation of an informationseeker’s problem” that involves “the limiting, labeling and framing of solution properties”(Marchionini, 1995). A search task is comprised of the information problem, itsmanifestation in the form of a query or series of queries, the search statements that are entered, and the items selected from a results list. In essence, the task is composed ofthe series of steps required to solve the information problem and that series of stepsvaries significantly with the type of system in use and the problem to be solved. Inrecent years, query and search statement have been used synonymously in particularwith studies of Web search queries (Jansen, Spink, Bateman &Saracevic, 1998). Thepractical implementation of such a study involves a researcher observing andcommunicating with a participant who is assigned one or more search tasks that he/shecompletes using a computer to access an information source.

In a review of selected IIRstudies published from 1982 to 1995, Yuan and Meadow (1999) classified the measuresfrom those studies into six groups that approximately parallel Tague-Sutcliffe’s (1992)variables. Typically, measures include:

a) Characteristics of the participant including biological, cognitive, socio-economic,and educational aspects;
b) Expertise, knowledge and expectations of the participant concerning the search topic, years of computer/searching experience;
c) Aspects of the search process, including decision-making, tactics, moves, errorsand so on;
d) Outcome measures concerning completeness of the task, satisfaction of the participant, relevance of the documents.

Often, multiple instances of datum are collected for items “b” to “d” depending on theexperimental variables that are manipulated, and/or the number of search tasksassigned.

A system designed to manage the experimental process must be able to handle theflow of all these elements by automatically manipulating variables and assigningtreatments to experimental units, as well as running a behind-the-scenes data collection mechanism to store data in a convenient format for later analysis.

To streamline IIR experimentation, while improving reliability and validity, we identifiedkey objectives for an experimentation system that could replace the traditionalapproach:

a) Significantly increase the size and heterogeneity of sample populations
a) Move the experiment out of the lab into a natural setting
b) Remove manual data entry by collecting data directly by computer to reducecosts and ensure data accuracy
c) Reduce (or eliminate) researcher-participant interaction to reduce demandcharacteristics and experimenter effects

III. Experiments On The Web – Alternatives To The Traditional Methods

Web-based experimentation has been an option for social science researchers since1995 when 'fill-in' forms were introduced in HTML 2.0. Since then increased technicalcapabilities have enabled the development of more sophisticated Web-basedexperiments (Krantz&Dalal, 2000). The Web venue was quickly adopted in the field ofpsychology for both research and teaching, as it allowed students to design as well asparticipate in experiments remotely. Today a number of sites support Web-basedpsychology experiments, including *The American Psychology Society(APS)* site (http://psych.hanover.edu/Research/exponnet.html), which contains links to over 100 Web-based survey and experimentation studies.

A typical Web-based experiment conducted to date includes a series of Web pages thatcontain the equivalent of a consent form, a survey to acquire personal characteristics ofthe participant and a page or more containing the stimulus. Anderhub, Muller andSchmidt (2001) investigated economic decision-making behavior using this process.They recruited participants via the Web using list-servers and responded automatically to an e-mail request with the URL of the experiment. First, participants were subjected to atest to determine eligibility to participate, e.g., German speaking with Java-enabledbrowser. They also added a set of criteria to validate participants: those using emailaliases and free e-mail accounts were eliminated, for example. Participants were paidfor their participation by bank transfer after having provided the necessary information inresponse to a questionnaire at the beginning. On receipt of the URL, participants hadone week to participate, and once started could neither interrupt nor stop the process;passwords were locked on activation. After each of the twelve decision points in thisexperiment, user-responses were submitted via a form to a database on the server. Thestudy used a NCSA web server and mySQL for database processing, and wasimplemented using Java applets. In studies of this sort, participants tend to read a pageor more of instructions and submit form-based data from a series of pages. Unlike IIR experiments, the experiment in these contexts tends to be self-contained within a set ofpages and the Web is used only as a conduit.

User interaction data

Another important way in which web search differs from traditional informationalretrieval is in the truly massive amount of user interaction data collectedby the major search engines. The most popular Web search engines are believedto log of the order of a billion interaction records each day.These interactionlogs include the queries people typed and the result links they clickedon. From them, search engines can assign general popularity ratings to pages(analogous to counting incoming links) (Culliss, 1999).

They can also use clicksto associate queries with pages and then use the associated queries in ranking(analogously to anchor text) (Xue et al., 2004; Hawking et al., 2006). Moredetailed interaction data collected by the user's own computer can be used toimprove search rankings (Agichtein et al., 2006).Click patterns can be used to deduce relationships between pairs of queriesand/or pairs of documents (Jones et al., 2006; &Szummer, 2007).Search engines may use click data as low cost relevance judgments for evaluatingand tuning their systems (Joachims, 2002; Joachims et al., 2005). Interactionsequences can also be used to suggest spelling corrections or related queries.Unfortunately, academic researchers have little access to search engine logsbecause of privacy concerns. Interaction sequences in logs may reveal a greatdeal of private information, even if the data contains no usernames or IP addresses. A good-faith attempt to make anonymous logs available to researchersin 2006, led to the unfortunate consequences described by Barbaro& Zeller(2006).

Spelling suggestions

Web search engines receive a significant number of misspelled queries. Someprovide a very helpful, did you mean X?" service. Due to the previously noted multi-lingual, neologism-prone characteristics of Web publishing, it is not at all feasible to make spelling suggestions by approximate searching within a normal dictionary word list. Nor is it useful to perform simple-minded approximatematching against the full vocabulary list of the Web as Web authors, like Websearchers, are highly prone to spelling errors. Very few misspellings would bedetected by this method and suggestions made could easily be foreign words orspelling errors.Details of commercial search engine spelling suggestion algorithms have notbeen published to our knowledge, but it is very likely that they are based onanalysis of query logs. Cucerzan& Brill (2004) describe and evaluate methodsfor spelling suggestions based on logs.

There are few more stages/issues which would include in web interactive information retrieval. Those are stemming, treatment of near duplicate content, SPAM rejection, adult content filtering, query targeted ads generation, snippet generation.

Web Search Strategies

- Analytical strategy (mostly querying)
  - Analyze the attributes of the information need and of the problem domain (mental model)
- Browsing
  - Follow leads by association (not much planning)
- Known site strategy
  - Based on previous searches
  - Indexes or starting points for browsing
- Similarity strategy
  - "more like this"

Non-search activities

- Reading and interpreting
- Annotating or summarizing
- Analysis
  - Finding trends
  - Making comparisons
  - Aggregating information
  - Identifying a critical subset

IV. CONCLUSION

Our electronic information world is becoming increasingly complex with more sources of information, types of information, and ways to access information than ever before. Anyone who searches for information is required to make more decisions about searching and expected to engage with an increased number and variety of search system. The Internet, in particular, has revolutionized the ability to search, especially in the commercial arena where we have the choice of using different search systems to search essentially the same electronic resources but with different interactive functionalities. The variability of data available, and the explicit or implicit structures of the data, also places a burden on both the searchers and system designers.

REFERENCES

1. AHLBERG, C., AND SHNEIDERMAN, B. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *Proceedings of the Human Factors in Computing Systems Conference (CHI'94)*.ACM Press, New York, (Boston, MA (April 24-28) 1994, pp. 313--317.
2. BELKIN, N. J., AND CROFT, W. B. Retrieval techniques.In *ARIST*, M. E. Williams, Ed. Elsevier, 1987, ch. 4, pp. 109--145.
3. EFTHIMIADIS, E. *Interactive Query expansion and Relevance Feedback for document Retrieval Systems*. PhD thesis, City University, London, UK, 1992.
4. HANCOCK-BEAULIEU, M., AND WALKER, S.An evaluation of automatic query expansion in an online library catalogue.*Journal of Documentation 48*, 4 (1992), 406--421.
5. HEWETT, T., AND SCOTT, S.The use of thinking-out-loud and protocol analysis in development of a process model of interactive database searching. In *Proceedings of INTERACT'87* (Amsterdam, 1987), Elsevier, pp. 51--56.
6. KOENEMANN, J., QUATRAIN, R., COOL, C., AND BELKIN, N. J. New tools and old habits: The interactive searching behavior of expert online searchers using inquery. In *TREC-3. Proceedings of the Third Text REtrieval Conference* (Washington, D.C., 1995), D. Harman, Ed., GPO, pp. 144--177
7. LANDAUER, T., EGAN, D., REMDE, J., LESK, M., LOCHBAUM, C., AND KETCHUM, D. Enhancing the usability of text through computer delivery. In *Hypertext: A psychological perspective*, C. McKnight, A. Dillon, and J. Richardson, Eds. Ellis Horwood, New York:, 1993, pp. 72--136.
8. Catledge, L.D. &Pitkow, J.E. (1995). Characterizing browsing strategies in the World-Wide-Web.*Computer Networks and ISDN Systems*, 27, 1065-1073.
9. Jansen, B. J., Spink, A., Bateman, J. &Saracevic, T. (1998). Real life information retrieval: A study of user queries on the Web. *SIGIR Forum,* 32(1), 5 -17.
10. Hert, C. and Marchionini, G. (1998). Information seeking behavior on statistical websites: theoretical and design considerations. In *Proceedings of the 61st ASIS Annual Meeting*, (pp. 303-314). Medford, NJ: Information Today.
11. Holscher, C. &Strube, G. (2000). Web search behavior of Internet experts and newbies. *Computer Networks* 33, 337-346.
12. Toms, E.G. (2000).Understanding and facilitating the browsing of electronic text. *International Journal of Human Computer Studies*, 53(3), 423-452.
13. Toms, E.G., Kopak, R., Bartlett, J., and Freund, L. (2001). Selecting versus describing: a preliminary analysis of the efficacy of categories in exploring the Web. In *Information Technology: the Tenth Text RetrievalConference, TREC 2001*, (pp. 653-662). Gaithersburg, VA: NIST.
14. Toms, E.G. d Freund, L., & Li, C. (2002).Building effective queries. In: *The 11th Text Retrieval Conference, TREC 2002, November 20-22, 2001, Gaithersburg, VA.* (in press)
15. Tague-Sutcliffe, J.M. (1992). The pragmatics of information retrieval experimentation, revisted. *Information Processing & Management* 28(4), 467-490.
16. Yuan, W. & Meadow, C.T. (1999). A study of the use of variables in information retrieval studies. *Journal of the American Society for Information Science*, 50(2), 140-150.
17. Anderhub, V., Mueller, R., & Schmidt, C. (2001).Design and evaluation of an economic experiment via the Internet.*Journal of Economic Behavior &Organization*, *46*, 227-247.
18. G. Culliss (1999). `User Popularity Ranked Search Engines'. Presentation at Infonortics Search Engines Meeting.
19. G.-R. Xue, et al. (2004).`Optimizing web search using web click-through data'. In Proceedings of ACM CIKM 2004, pp. 118.
20. D. Hawking & N. Craswell (2005). `Very Large Scale Retrieval and Web Search'.
21. E. Voorhees & D. Harman (eds.), TREC: Experiment and Evaluation in Information Retrieval, pp. 199-232.MIT Press
22. E. Agichtein, et al. (2006). `Improving web search ranking by incorporating user behavior information'. In Proceedings of ACM SIGIR 2006, pp. 19-26, NewYork, NY, USA. ACM Press.
23. N. Craswell& M. Szummer (2007). `Random walks on the click graph'. In Proceedings of ACM SIGIR 2007, pp. 239- 246.
24. R. Jones, et al. (2006). `Generating Query Substitutions'. In Proceedings of WWW 2006, pp. 387-396, Edinburgh, Scotland.ACM Press.
25. T. Joachims (2002). `Optimizing search engines using clickthrough data'. In Proceedings of ACM KDD 2002, pp. 133-142.
26. T. Joachims, et al. (2005). `Accurately interpreting clickthrough data as implicit feedback'. In Proceedings of ACM SIGIR 2005, pp. 154-161.
27. M. Barbaro& T. Zeller, Jr. (2006). `A Face Is Exposed for AOL Searcher No. 4417749'. The New York Times http://www.nytimes.com/2006/08/09/technology/09aol.html?ex=1312776000.
28. S. Cucerzan& E. Brill (2004). `Spelling correction as an iterative process that exploits the collective knowledge of web users'. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 293-300, Barcelona, Spain.Association for Computational Linguistics.