Partnerships in Navigation: An Information Retrieval Research Agenda

Michael Buckland
Mark H. Butler
Youngin Kim
Barbara Norgard
Christian Plaunt

School Information Management and Systems
University of California
Berkeley, CA 94720-4600
LP01

In: Forging New Partnerships in Information: Proceedings of the 58th Annual ASIS Meeting, 1995, pp. 84-89. Medford, NJ: Information Today, 1995

Abstract

The transition from searching in a single database to searching a multiplicity of networked atabases exacerbates some old difficulties in the design and evaluation of retrieval systems and creates new ones. A network environment calls into question the traditional definitions of recall and relevance. Efficient network searching raises questions about where to look first, where to look next and when to stop searching. The need for "entry vocabulary" support and the need for support in moving from one system vocabulary to another are increased by the increased use of more different databases. The network environment offers the option of collecting different representations of the same object and merging them into an extended record.

1 Introduction

Information retrieval theory and practice developed in relation to searching in a single database. Each database had a single search engine. The 1995 environment of large numbers of network accessible databases is quite different. Both the theory and the techniques of information retrieval need systematic revision to support partnerships in navigation.

The paper has two components:

- Problems: A summary of necessary steps in information retrieval design required for the transition from traditional "single database" retrieval to the emerging "networked resources" retrieval.

- Solutions: A presentation of research and experimental solutions recently completed at Berkeley, addressing what is needed to move information retrieval techniques from a single database environment to a network environment.

2 Problems

1. Recall and relevance. The traditional definition of recall is based on the number of relevant items in a single database. But this basis ceases to be meaningful in the new context of navigating endless numbers of networked databases. The definitions of recall and precision need to be revised in order to be suitable for network as well as single database environments.

2. Network navigation. Deciding which databases to search was not much of an issue in traditional information retrieval. Absent a "catalog of the internet", aids are needed for the three most basic decisions of network searching: Where to look first? Where to look next? and, When to stop searching?

3. Entry vocabulary and cross vocabulary assistance. In traditional retrieval the system vocabulary of the database would be more or less familiar. When navigating in a network environment of multiple heterogeneous databases, such familiarity is much less likely. An effective and efficient "entry vocabulary" function mapping natural language queries onto system vocabularies becomes even more important.

An important feature of networked resources is the ability to extend searches from one database to another. This implies extending a search query from one system vocabulary to another. A "cross vocabulary" function is needed for associating terms from one domain vocabulary with terms of another domain vocabulary.

4. Records. In a single database the retrieved records are of a predetermined content. The only option is how much of it to display. Since networked databases often contain varied representations of the same object, by extending the search integrating information cumulatively an "extended object" can be built up.

3 Solutions

3.1 Recall and Relevance

The traditional definition of recall is based on the number of relevant items in a single database. The familiar, serious difficulties with this approach are greatly exacerbated in the transition to the navigation of networked resources:

(i) It is notoriously difficult to estimate how many relevant items there are in a database. The difficulty explodes exponentially as the number of networked databases increases. Given the diversity of the contents and vocabularies of the ever increasing number of networked databases, the idea of determining what would constitute 100% recall in a network environment becomes less and less realistic.

(ii) Even when the search is only in a single database, it is known that users are not usually interested in all relevant items, but in obtaining a select few instances (Larson 1991). This seems to indicate that the traditional definition of recall is inappropriate for most searches. Even if a searcher were interested in all instances of relevant documents, the realism of this objective begins to sag when the searcher is faced with the prospect of navigating an indefinite, unstable, and potentially endless array of databases. The definition of recall needs to be revised to apply to network as well as single database environments.

We have chosen to adopt an alternative definition of 100% recall, one closer to searchers' observed behavior: 100% recall is reached when the number of relevant items desired by the searcher has been retrieved. To retrieve more relevant items than is wanted is superfluous (at best) and cannot be regarded as any kind of improvement in retrieval performance. When a searcher wants one instance of relevant items (or a few), retrieving one instance (or a few) efficiently should be the objective of the search and, therefore, the basis for assessing retrieval performance. This more generalized definition is inclusive: It allows for the special case of when the searcher does want to retrieve all relevant items when that is a realistic option. A pragmatic definition of "demanded recall" as the percentage of the number of relevant items sought is, therefore, a more general definition than the traditional definition which is limited to the special case of all relevant items.

(iii) Relevance is an elusive concept. Assessing satisfaction with retrieval performance is problematic because the notion of relevance is rooted in subjective and unstable perceptions of an ill-defined quality (Schamber 1994). No amount of statistical artistry can overcome the fundamental impossibility of consistent, measured evaluation based on subjective perceptions of an undefined and intangible essence. This is not at all to suggest that traditional approaches to evaluating relevance should be abandoned, but, rather, to acknowledge the incorrigible difficulty involved and to suggest that, at times, an alternative definition might have its uses. The alternative that we have been using is more pragmatic, but not necessarily better: Relevance is defined operationally by the presence (or absence) of attributes specified by the searcher. This "pragmatic relevance" need not be a binary condition: An item with more of the attributes specified by the searcher is more relevant (on this definition) than one with fewer. We have developed different search techniques based on these alternative assumptions of recall and relevance in a network environment which will be discussed below.

3.2 Network Navigation

The OASIS research program (Buckland *et al.* 1993) uses the University of California's MELVYL system for experiments in network navigation in more than one way:

(i) MELVYL provides access to multiple databases, so it can be used for experiments in extending searches among heterogeneous databases.

(ii) The OASIS front-end to MELVYL can telnet to other network accessible databases. In particular, as part of the CNRI Computer Science Technical Reports project we query Computer Science Technical Reports servers on the World Wide Web.

(iii) MELVYL contains records drawn from the 100 libraries of the nine campuses of the University of California (UC). The name of the holding library and campus is retrieved for every record retrieved. It is therefore possible in post-processing to rearrange any retrieved set by the libraries (or campuses) that hold the retrieved items. This rearrangement by location, in any desired order, presents what would have been found if one had searched these locations separately, serially, in that order. Thus MELVYL can be used as a virtual network to simulate searching in a realistic network of up to 100 nodes without the added worry of compatibility problems.

We are concerned with two aspects of network searching: efficiency and adaptiveness. Efficient searches travel to the smallest number of nodes necessary to answer a query. Adaptive searches take advantage of user preferences to modify searches in response to evolving retrieval results. Adaptive searches are conditional search commands that can be altered throughout the search process.

```
CAT-> check all DBMS
Searching CAT...The catalog of books in the UC libraries
    CAT contains 31 records
Searching CC...A database of 6,500 scholarly journals
    CC contains 37 records
Searching COMP...A database of 200 computer-related journals
    COMP contains 13343 records
Searching INS...A database of 4,000 physics, electronics and
computer journals
    INS contains 7498 records
Searching MAGS.....A database of 1,500 magazines and journals
    MAGS contains 2703 records
Searching NEWS...A database of 5 major U.S. newspapers
    NEWS contains 89 records
Searching PSYC...A database of materials indexed by the
American Psychological Association from 1967 to present.
    PSYC contains 2 records
Searching ABI...A database of business and management journals
    ABI/Inform contains 3934 records
Searching CSTR...A network of servers with CS Tech Reports
from UCB, CMU, MIT, Stanford, Cornell, Dartmouth and more...
    CS-TR contains 39 records
```

Figure 1: Example of the CHECK command probing multiple databases for the keyword "DBMS"

3.2.1 Efficient Network Searching

Efficient network searches present three problems: knowing which databases to search, the order to search them in, and when to stop searching. As the network is navigated the costs of continued searching are increasingly likely to outweigh the remaining benefits to be gained.

OASIS implements one solution to the problem of knowing which databases to search with a "CHECK" command. Given a query, OASIS will simultaneously search ("CHECK") each of the databases on MELVYL and the Computer Science and Technical Reports database via the World Wide Web. Figure 1 shows a simultaneous probe of several databases in a search for the term "DBMS" in any keyword field (the title, subject, abstract fields). This approach works well against a small number of databases, but does not scale well.

In addition to knowing which databases to search, it is important to know when to quit searching. Visiting every database in a networked environment should be avoided whenever feasible because the principal result will be increased retrieval of duplicative and non-relevant materials at great expense. An example of this is illustrated by a search for the subject heading "ontology". Records matching this query are to be found at 15 locations in the UC system according to MELVYL. When arranged by the number of volumes at each site, the five sites with the most material each hold 25-38% of the titles in the UC system for this query. There is clearly substantial overlap in the titles held at these locations. See Figure 2.

Site:	UCLA	UCB	UCSD	NRLF	UCI
% of total	39%	33%	30%	26%	25%

Figure 2: Example of distribution of material at each location for a search on the subject heading "ontology"

Site:	UCLA	UCB	UCSD	NRLF	UCI	Other
% of total						
unique items	39%	16%	7%	15%	3%	20%
cumulative %	39%	55%	62%	77%	80%	100%

Figure 3: Example of distribution of unique items at each location for a search on the subject heading "ontology"

When viewed in terms of unique volumes a very different picture emerges. After the first site, searching each subsequent site contributes less and less benefit while maintaining constant the cost of visiting each site. UC Irvine has 25% of the objects matching a query for the subject heading "ontology", but only contributes 3% of the unique objects matching the query when it is visited after the four other sites. See Figure 3.

Return on the investment in extending a search around such a network clearly diminishes rapidly. In this case the first two sites contain over half of the unique items that would be retrieved by searching the entire network. The first four sites contain over three-quarters. Except for the untypical case in which the searcher does want to retrieve all instances, visiting each of the remaining 11 sites is unlikely to be worthwhile.

This example also illustrates how database size can be a simple yet effective factor in deciding which database to search next. The CHECK command provides information on the number of items that can be retrieved from a database which allows OASIS to visit the databases in size order or for the searcher to modify the search in light of amount and type of material to be found in each database (See (Gravano *et al.* 1994) for related work on this topic).

3.2.2 Adaptive Network Searching

Adaptive searching takes advantage of the conditional nature of searches. One typical condition (often implied) involves the number of items a searcher wishes to find. There are three basic variations on this theme.

(i) The first is find any instance: any one example with the specified attribute value(s) will suffice. "Known item" searching is a very common example.

(ii) The second type of search involves finding some instances of an item, which requires searching until an acceptable number of occurrences have been discovered or until all the potential sources have been searched, whichever occurs sooner.

(iii) The third type of search involves finding every instance. This is the case upon which information retrieval research has been predicated almost exclusively, even though it is not common

```
Search request: F TW ROMEO JULIET AND LANG ENGLISH
Search result:  383 records at all libraries

CAT-> summ test dt

229 BOOK
52  SOUND RECORDING
49  MUSIC SCORE
36  VIDEORECORDING
10  DISSERTATION
4   MOTION PICTURE
2   VIDEORECORDING, DISSERTATION
1   ANALYTIC
```

Figure 4: An example of a large retrieved set sorted by document type from which the user could select just one or a few instances of media

in practice.

Two other types of searches are the extreme instance and the preferred instance. An example of the extreme instance search would be finding the most recent review article on a topic, the earliest book or the latest version. In such a case, all potential sources must be searched because the last database might contain the most extreme case. The preferred instance search is the weaker form, where one (or more) of the specified attribute-values are desired but not absolutely insisted upon.

Preferred instance searches have interesting attributes. Preferences can vary in strength from required to inconsequential. The traditional query "find subject ontology" specifies that the attribute-value pair subject-ontology is required and that there is complete indifference with respect to all other attribute-value pairs. This is, of course, a crude, implausible, but convenient simplification. The strength of preferences can be expressed on a scale rather than in a binary form as has been done in some extended Boolean systems.

Preferences can also be conditional. For example, the user may prefer texts in the English language but, if they are absent or too few, will accept texts in French and Spanish, especially if they are recent.

In the last resort, when a search has been reduced to a mere handful, some variety with respect to one attribute may be desired. For example, given a search for material on a particular disease, the final step might be to weed the retrieved set in such a way as to ensure that the final few contain a variety of different medical treatments. Similarly, one might want to sort the retrieved set by document type as a basis for selecting one or more of each type.

In another example, faced with a large retrieved set of materials about "Romeo and Juliet", the option of selecting just one instance (or a few) of each of the diverse media would be attractive. See Figure 4. We are implementing a "VARY" command which allows the user to select a variety of values for a particular field or a bibliographic record from the sorted set.

3.3 Entry Vocabulary And Cross Vocabulary Assistance

We identified the need for support for (i) entry vocabulary (i.e. for translating the searcher's terminology into the system's terminology) and (ii) for mapping from one system vocabulary into another. The former can be treated as a special case of the latter.

In our preliminary work we have been testing a two stage algorithm which establishes a probabilistic relationship between (a) title words, author names, and abstract words and (b) subject headings in a controlled vocabulary. Given any combination of title words, author names and abstract words, what would be the most suitable subject headings?

Here we are exploiting the knowledge in catalog records to which subject headings have been assigned by human indexers for the INSPEC database, one of the databases provided by MELVYL. By observing the associations between lexical clues found in titles, authors and abstracts of a large set of human indexed catalog records, this algorithm can be trained to predict appropriate subject headings for new titles (and abstracts) when they are presented as queries to our system. We are using a "collocation" technique to assign subject headings to documents based on data gathered from bibliographic records. The training phase identifies and extracts content-bearing lexical items from elements found in a large set of existing bibliographic records (authors, titles, subjects, abstracts, etc) and "collocates" them with manually-assigned subject headings (controlled vocabulary index terms) to form a "dictionary".

We interpret collocation in a broad sense to mean that there is some measurable association between the extracted lexical units and the assigned indexing units. Using these associations, we create a dictionary of mapping to use in the deployment phase. The dictionary can be tested using other existing records not used in the training set. As these test documents are presented to the system, each lexical unit's associations are looked up in the dictionary and put into a ranked list. These ranked lists can be consulted by the user to predict likely subject heading assignments for each document.

Figure 5 shows an example of a search against the dictionary. Given the title of a document ("Rethinking library education in the information age"), the system generates a ranked list of subject headings. A "T" indicates that a subject had been correctly predicted. "NIL" indicates a subject heading that been wrongly predicted. Even though the headings marked "NIL" indicate that indexers might not have assigned them to documents with these terms in their titles, majority of them are quite obviously related.

There are several uses for such a system:

(i) To prompt the searcher with likely subject headings when a search is extended into another database.

(ii) Since title and abstract words can be used as an approximation of the searcher's natural language, such a system can also be used as an entry vocabulary in an initial search.

(iii) For computer-assisted subject heading assignment and computer-assisted quality control of existing indexing.

These kinds of operations are becoming increasingly important especially in a widely distributed and uncontrolled internet environment.

3.4 Cumulating Information in a Network Search: Extended Records

A predictable consequence of navigating a network is that different but related records will be found in different databases, e.g. different bibliographic records representing the same document. For a more detailed discussion, see (Buckland *et al.* 1994).

```
Words: RETHINKING LIBRARY EDUCATION IN THE INFORMATION AGE

  1  LIBRARY-AUTOMATION: T
  2  INFORMATION-SCIENCE: T
  3  EDUCATION: T
  4  INFORMATION-SERVICES: NIL
  5  INFORMATION-RETRIEVAL-SYSTEMS: NIL
  6  LIBRARIES: NIL
  7  INFORMATION-CENTRES: NIL
  8  CATALOGUING: NIL
  9  MANAGEMENT-INFORMATION-SYSTEMS: NIL
 10  EDUCATIONAL-COURSES: NIL
```

Figure 5: A test: Three assigned subject headings were correctly predicted. No assigned subject headings were missed. Seven subject headings that were predicted had not been assigned

The example in Figure 6 illustrates how records found in different sources, but representing the same objects, can be amalgamated to form an extended representation of the object. Beginning with a record from the Computer Science and Technical Reports database, OASIS will automatically query MELVYL to find records containing additional information about the same object.

Two purposes are served with such a command. The first is to provide a history of the object. This command can be used to find versions of a technical report that are subsequently published in journals. There may or may not be content differences between the instances represented by the two records. The second purpose is to provide additional information such as subject headings or additional abstracts that have been assigned in one database but not another.

4 Summary

Information retrieval in a network environment exacerbates some old problems and creates some new ones. We suggest some solutions including pragmatic interpretations of recall and of relevance, an experimental simulation of the network environment, a technique for extending a search to unfamiliar system vocabularies, and methods for supporting decisions concerning where to search and when to stop searching.

Acknowledgements

This paper reports on work done as part of the OASIS research program led by Michael Buckland. It was supported in part by the US Department of Education HEA IID R197D40008, “Online access in multiple database environments” and by the Computer Science Technical Reports Project funded by the Corporation for National Research Initiatives.

<AFFILIATION:INS> California Univ., Berkeley, CA, USA
<ABSTRACT:CSTR> Generic design articulated through Information Processing Theory.
<ABSTRACT:INS> The notion of generic design, although it has been around for 25 years, is not often articulated; such is especially true within Newell and Simon's (1972) information-processing theory (IPT) framework. Design is merely lumped in with other forms of problem-solving activity. Intuitively, one feels there should be a level of description of the phenomenon that refines this broad classification by further distinguishing between design and nondesign problem solving. However, IPT does not facilitate such problem classification. The authors make a preliminary attempt to differentiate design problem solving from nondesign problem solving by identifying major invariants in the design problem space.
<TITLE:CSTR> Motivating the notion of generic design within information processing theory: The design problem space.
<TITLE:INS> Motivating the notion of generic design within information-processing theory: the design problem space.

Figure 6: Example of extended record integrating information from 2 databases (INSPEC and CSTR). This record shows only INSPEC has affiliation information. Both have abstracts, but they are different

References

Buckland, M. K., M. H. Butler, B. A. Norgard, & C. Plaunt (1993). Oasis: A front-end for prototyping catalog enhancements. *Library Hi Tech*, 10(4):7–22.

Buckland, M. K., M. H. Butler, B. A. Norgard, & C. Plaunt (1994). Union records and dossiers: Extended bibliographic information objects. In *Navigating the Networks: Proceedings of the ASIS Mid-Year meeting, Portland, 1994*, pages 43–57, Medford, NJ. Learned.

Gravano, L., H. Garcia-Molina, & A. Thomas (1994). The effectiveness of GLOSS for the text database discovery problem. *SIGMOD Record*, 23:126–137.

Larson, R. R. (1991). Between Scylla and Charybdis: Subject searching in the online catalog. *Advances in Librarianship*, 15:175–236.

Schamber, L. (1994). Relevance and information behavior. In M. E. Williams, editor, *Annual Review of Information Science and Technology*, volume 29, pages 3–48. Learned Information, Medford, NJ.