Supporting Consultants with Task-Specific Information Retrieval

Kurt D. Fenstermacher and Cameron Marlow
University of Chicago
Department of Computer Science
1100 E 58th Street
Chicago, IL
{fensterm, camarlow}@cs.uchicago.edu

Abstract

Although information retrieval systems based on the words and phrases in documents offer excellent performance, such systems are limited by their lack of knowledge about documents. When the goal is the management of knowledge to support reuse, information systems must be context aware. The information need that leads to a search for documents has a context, as do the documents of the search space. Only by matching the contexts of the documents, as well as the words and structure of the documents, can knowledge be effectively reused. We describe the SINGLESOURCE system, which has been designed to facilitate knowledge reuse within a group of professional researchers at a major consulting firm.

1. Introduction

Most information retrieval systems represent documents as combinations of words and phrases, and build indexes to track word and phrase counts within and among documents. (Salton and McGill 1983) Using powerful statistical techniques, modern retrieval systems offer a variety of powerful search techniques. For example, the Verity engine allows for proximity searching, word stemming, field-specific searches, dynamic clustering and more. (See (Salton and Yang 1975) for a description of vector-based techniques and (Verity 1999) for more information on the Verity Developer’s Kit.) Despite these powerful features, many people find it difficult to locate appropriate documents using today’s search tools.

The enormous popularity of the World Wide Web has been accompanied by the frustration of finding “just the right” document to address an information need. Despite the more constrained environment, often intranet search engines suffer from the same problems. For example, imagine a sales representative about to make a sales call on a prospective customer searches a corporate database for product descriptions in the company’s new product line. After entering the query “Xifor 5000 features”, the salesperson hastily prints out the report returned as the second search result and heads for the customer’s headquarters. In the middle of the sales call, however, the sales representative discovers that the report is actually a critique of the current Xifor 5000, written by the engineering department to stimulate discussion of product improvements. Rather than hand over a report that highlights the product’s deficiencies, the salesperson calmly turns the conversation to the unseasonably warm weather.

The problem is that the corporate search engine failed to recognize the document context. The sales representative of the above example needed a discussion of product features that would be appropriate to a sales call, not an engineering brainstorming discussion. The word-based retrieval of the company’s document search can’t distinguish between two documents that contain similar words and phrases, and yet were authored in (and for) different contexts. Recognizing the importance of context, however, is far from recognizing the context itself. How are we to build systems that support context-assisted retrieval?

Before context can be used in retrieval of documents of course, the context must be associated with documents. There are two means of developing the context of a document: applying natural language processing (NLP) techniques to glean the context of a document from a close reading, or asking document authors to annotate their work. Although NLP systems have made great strides over the past decades, deriving the context of a document is likely to be an “NLP-complete” problem — automatic context deduction is not feasible with the technology of the late 1990’s. Unfortunately, few people enjoy annotating their work. Programmers (especially undergraduate ones) resist commenting their code, sales representatives avoid write-ups of sales meetings and physicians hurriedly dictate notes as they move from one patient to the next. The challenge for information systems designers is to extend the reach of computer information processing, while using the expertise of people to aid in the most difficult cases.

The goal of adapting information retrieval technology to aid in the reuse and capture of relevant knowledge has become a much-bandied buzzword: knowledge management. Over the years, the emphasis has moved from storing and accessing data to retrieving information and now managing knowledge. The progression of popular terms carries a grain of truth with it: managing knowledge is distinct from retrieving information. Effective knowledge management requires a richer representation of documents and their components than the vector space models popular in information retrieval. Along with a deeper representation, knowledge
management demands more sophisticated techniques for manipulating that representation.

Our approach, used in the SINGLESOURCE system we have built, is to construct task-specific tools that are context-aware. The tools are designed to facilitate different phases of complex tasks and to work together. Each tool annotates its work to reflect the context and is aware of the context created by the previous tool in the chain. Together, the tools create a working environment tailored for programmers, sales representatives, physicians and others. In the next section, we outline the original motivation for SINGLESOURCE, and some of the research goals for the system.

2. The Problem

One of our lab’s funding partners, a major consulting firm, asked us to evaluate opportunities for streamlining the firm’s internal research process. Within the firm, an internal research group (hereafter called Research, or “R&I”) provides reference library and search services to the firm’s consultants in the field. For example, a consultant who is studying the impact of consolidation in the United States telecommunications industry might request an overview of the industry, including profit and loss and market share figures for the top twenty national firms.

Our first task was to analyze the process for handling requests for information (often shortened to simply “requests”). Consultants pose requests by phone, fax or email. (Walkup requests are rare, but do occur.) The request is made through a central station, and the researcher on duty enters information about the request into a database, called the request tracking database. The researcher then tries to categorize the request according to the type of research needed (we have labeled this step “request triage”). Occasionally, more complex requests require discussion before they can be assigned to a researcher. The discussion is either done by phone and voice mail, or in a regular morning meeting if more involved discussions are needed. The morning meeting usually took place at 9 a.m., and so requests were sometimes deferred until “the 9 o’clock meeting”. One of our research goals was to implement a system that made the nine o’clock meeting unnecessary, through automated request assignment.

Simple “lookup” questions are often handled within minutes, often by the researcher staffing the desk. A typical “lookup” question is “What is the GDP per capita in India, for each of the past 5 years?” More difficult requests require document location, analysis and synthesis, e.g. a request to identify the five most critical factors in determining the success of domestic mutual funds. Some requests require specialized expertise — the general researchers refer these to area experts within the firm’s "practices". Describing the impact of the deregulation of the United Kingdom telecommunications market on British Telecommunications (BT) falls to specialists in European telecommunications.

After a request had been assigned, the researcher would sometimes contact the requesting consultant for clarification and narrowing of the scope. Frequently, the problem was that the requesting consultant, who was intimately familiar with the context of the question, would pose a request that would be much too broad when taken out of context. (For example, “Send me all the information you have on the Bell operating companies since the breakup of AT&T.”) The researcher would discuss the motivation for a request and the project context, which would then narrow the question much further. Encoding the context of a request along with the request was another of our research goals.

2.1. Information access

Information was accessed in several different ways, including the firm’s proprietary databases, subscription information services, public information sources. For lookup questions, the goals are speed and accuracy. (Speed can be important because a consultant might have stepped out of a client meeting to ask for information to respond to a client question. The answer might be needed before the consultant returns to the meeting.) Accuracy concerns are addressed by selecting trustworthy sources, and cross-checking information. The key to speed is knowing where to look. One of our goals was to develop strategies to improve the speed of lookup responses by capturing search strategies for reuse.

More difficult requests involve more involved search and analysis. For the more open-ended questions, the problem is that relevant information is rarely already packaged in a single source. Thus, identifying relevant documents is more challenging, and those documents must be further analyzed, often being summarized in a cover memo sent to the requesting consultant. The researcher then enters any remaining information about the request or response into the request tracking database, completing the process.

One of the challenges facing researchers is the diversity of sources and the lack of integration. To search multiple sources, the researcher must often switch from one search tool to another (Fenstermacher and Hammond 1998). New search tools mean new interfaces, different query languages — each similar enough that they look alike, and yet different enough that researchers must be familiar with all. The many different result formats used by different search tools makes it difficult to consider a set of the best documents across sources. The firm has addressed this issue in part by creating an all-in-one search that scans several of the firm’s proprietary databases and returns a list of the number of matching documents within each database. However, the system did not give researchers any information regarding the content or relevance of the returned documents, and so offered no benefit.

The display of search results was also problematic. As professional researchers, the group we worked with
wished to make the relevance judgment themselves. In other words, they wished to sacrifice precision for recall. Often researchers would browse broad document sets, choosing a few to focus on, and then explore those chosen in more depth. One of our research goals was to make searching smarter — both by offering a single query interface for multiple sources, and by integrated display of the results.

Different kinds of requests call for different search strategies. With lookup requests, the goal is clear: find the specific datum that answers the question. The specificity of the request can make it difficult to locate the needed information, however. For example, if the request refers to GDP (the gross domestic product), then GNP (gross national product) figures will be incorrect. A researcher who remembers finding GNP data, but mistakenly recalls it as GDP will waste time (and possibly money) tracking down an incorrect answer. In our research, the strategy for accessing information is an important element in handling requests. Noting the strategy of access is more general than noting the specific information retrieved, and thus more likely to be reused in the future (either by other researchers assisted by the system, or the system itself).

Efficiency of information access is critical to the research group we worked with. Request handling must not only be quick and accurate, but also as inexpensive as possible. Some of the information sources used by the researchers have high connect fees and document retrieval charges. The costs of research include not only the cost off access, but the researchers' time as well. If less-skilled individuals can handle a request there will be cost savings from the use of the system.

2.2. Reuse of knowledge

When asked if the same requests occur frequently, most researchers replied that, no, every request was different. When the specifics of each request are considered, few requests reoccur. However, when the focus is shifted to a broader perspective, the same request patterns do tend to recur. The recurrence of requests, coupled with the knowledge intensive nature of request handling suggested an opportunity to reuse knowledge, leveraging the group’s experience. The focus on reuse was on two aspects: knowing where to look for needed information, and given that, knowing how to look for it.

Often the challenge in handling requests lies in identifying the appropriate sources to search. With so many information sources available within the research group, the challenge was often to select the appropriate source. Since no meta-index of sources and their strengths existed, researcher would need to scour the available sources considering dozens of candidate sources. Often a researcher who had handled a previous request for similar information had identified an ideal source, for example a contact at a trade association, but the contact and its relevance was not recorded. Thus, the current researcher would need to “rediscover” the source.

If the relevant source were a human expert, then the researcher could simply contact the expert and pose the question. When using online sources, however, researchers need to know more than simply where to find the source. The search strategy used to locate the information would be critical as well.

2.3. Information profiles

One of the burdensome aspects of request handling is the process of managing the flow of requests. When a request is initially submitted, the request needs to be assigned to an individual for handling. Throughout the process, information about the request is being entered and queried. When a request is received, administrative information, including the requesting consultant, budget, deadlines and other information is entered.

3. Design criteria

At the heart of the SINGLESOURCE system is the goal of improving the quality of knowledge reuse by context-aware retrieval. At each step of the request handling process, the researcher handling a request should be offered helpful, relevant assistance. After a request is routed to a researcher (possibly by SINGLESOURCE itself), logging in to the system will display not only the newly assigned request, but also knowledge needed to handle the request. This knowledge would likely include the context of the consultant’s information need, results of automatically generated searches and pointers to people whose expertise would be helpful. In this expertise can be of two kinds: research expertise (knowing how to find relevant information effectively) and expertise of knowledge (actually knowing the relevant information). Although not all these capabilities currently exist, some do and the framework has been built for others.

3.1. Context maintenance

In the course of developing the system, we quickly realized that at each stage of the request handling process, the system needed the context of the task to intelligently assist the user. Furthermore, the context of one phase of the task was inherited from the previous phase. Thus, SINGLESOURCE is designed as a single working environment in which all of the tools share a single representation of the task context. Currently, the shared data for the task context is stored in a relational database accessed by all SINGLESOURCE applications, and the data is used to build the context for each task. Our current research is refining this representation, with plans to migrate the system to an object database that allows more complex searches based on the relationships among objects (consultants, researchers and documents).

Another goal of the SINGLESOURCE design is to integrate information filtering using dynamically built profiles of system users. Information profiling, as discussed in (Belkin and Croft 1992), is not a replacement...
for information retrieval, but a complement to it. The goal in SINGLESOURCE is to represent users’ interests and expertise to facilitate information access and sharing. By representing user interests, the system can filter information and highlight topics for particular users. The same profile can also be used to identify users’ competencies. For example, a researcher who frequently uses search tools that offer automotive industry information is a likely candidate to handle a recently submitted automotive request.

4. SINGLESOURCE implementation

4.1. Consultants

The entire process of information retrieval begins when a consultant decides that some sort of information would complement his work. SINGLESOURCE provides the consultant with an application intended to satisfy all of his information needs. The interface to this application separates the consultant’s information needs into two categories: standard and custom requests. In the initial stages of the information request process, it may be the case that the consultant’s questions can be answered by means of standard information types. These types are broad classes of documents that are frequently referred to in a consultant’s job, and have been provided in a standard interface for easy retrieval (Figure 1, left side). In addition, a consultant may perform elementary Internet or Intranet searches which may answer their information needs. In the event that a consultant’s information needs have not been met by any standard request, they can consult the services of R & I via the custom request interface (Figure 1, right side). The information that the consultant enters in this interface will serve to identify the request and to provide the researchers with the necessary information to complete their work. In comparison to a normal query, such as in the standard interface, this query has associated with it a particular context: the consultant, the type of request, the consultant’s client and the request itself.

Consider, for example, a consultant who is working with Berkshire Hathaway, a holding company that owns several insurance interests. She is currently interested in the opportunity for Berkshire Hathaway to purchase a reinsurance company, and decides to consult SINGLESOURCE. After reading a financial summary for Berkshire Hathaway and attempting a few Internet searches of her own, she decides that she should submit a request. This task is explicitly the process of collecting documents and commenting on them. For this reason, we liken the request for research to a clipboard, to which documents and the researcher’s observations will eventually be attached.

4.2. Researchers

When a request for research is placed, it is saved to a networked database, and is immediately available to the research (R & I) department. SINGLESOURCE provides a tool for the management of research requests within R & I. It can be used at meetings to discuss and assign requests, and for administrators of R & I to manage the workload in their departments. Since researchers can see all of the clipboards at once, and have tools to rearrange this information, the process of research is made more efficient. Similar requests can be grouped, workloads can be managed, and unassigned or uncompleted requests are quite visible. In our particular example, at a department-wide meeting the request for information on Berkshire Hathaway is assigned to a particular researcher (Figure 2).

**Figure 1** – Consultants can request standard information packages or request their own

```plaintext
Choose a standard product:

R & I Product Menu:
- Intranet search
- Company financial summary
- Order a PONet Document
- Find an expert
- Search the Web
- Company/Industry news

Global Intranet Search

Please enter your search terms:
- strategic alliances insurance reinsurance operations

Search Global Intranet

Enter your request:

Respondent: [mm/dd/yy]
Client: Berkshire Hathaway
Change Code: BH-0787

Request:
- Synergistic opportunities for BH and a reinsurance company

Date of Request:
- December 12, 1998
Budget LPS: 14000

Request:
- Looking for information on strategic alliances between insurance and reinsurance operations

Clear
Save
```

Standard Request Custom Request Current Requests

Standard Request Custom Request Current Requests

Standard Request Custom Request Current Requests

Standard Request Custom Request Current Requests

Standard Request Custom Request Current Requests
The final piece of the SINGLESOURCE suite of tools is an application designed for use by the researcher; for a given request, all of the actual research occurs within this interface. Once a request has been assigned to a researcher, it immediately appears in a list of tasks on the researcher's agenda. As with the R & I task management tool, the researcher is able to view his tasks several ways to suit his preferences. Once he has decided to work on a particular clipboard, he is able to view all of the particular information related to that clipboard. This set of information establishes the context of the search, and this context will be the same as the one that the system uses when it aids research.

After selecting a clipboard, the researcher then begins the task of satisfying the information needs of the consultant's request. To aid in this process, SINGLESOURCE provides an interface consistent with the rest of its components. Nearly every on-line resource that a researcher may consult is provided, and the interface for interaction with each of these resources is removed. Similarly, the results from each of these resources are made to be consistent; the result is a uniform interface for searching multiple Internet resources simultaneously. The researcher merely enters his search terms, chooses his search tools based on the Internet resources offered (Figure 3), and then executes his search. Throughout the process, he is never asked to conform to the specifics of any particular search engine.

**Figure 2** – Requests are filtered to researchers via a standard interface

**Figure 3** – Researchers have their choice of search tools
Using the context of the search, along with an engineered taxonomy related to the tasks of a consulting firm, we categorize each of the returned documents into conceptual classes. With a hierarchy representation of these classes, the researcher is able to quickly scroll down to his desired topic and quickly eliminate many documents which have a completely separate conceptual meaning. They can then access the document by merely asking the system to display it (Figure 4). If they find the document relevant to the research topic, they can add it to the clipboard with their comments. In our example, the given researcher is looking for a document related to Berkshire Hathaway's stock stability; he searches for the terms “Berkshire”, “Hathaway”, “stock” and “insurance”. By scrolling down the conceptual hierarchy, he quickly locates the “stock” category and finds a document contained within regarding Berkshire Hathaway’s stock. After finding a passage negating the potential merger of Berkshire Hathaway, he adds the document to the clipboard with his comments about the passage.

When the research on a request has been completed, the researcher can automatically generate a report outlining all of his research and the documents which he discovered, which can then be e-mailed or sent back to the consultant. This completes the loop of the request, and it is removed from the list of active requests. Our research may be reused in the case of another question about Berkshire Hathaway or even in another request for information about another insurance company. This is possible only because of the level of integration within the SINGLESOURCE system.

5. Conclusions

The SINGLESOURCE system represents a new hybrid of information retrieval systems: one that builds on the success of word and phrase-based statistical IR systems, and combines it with lessons of context and representation gleaned from the study of natural language. We believe that this combination is the key to transforming information retrieval into knowledge management.

References