Chapter 9
Collaborative Retrieval Systems
Reusable Information Quests

Ying Sun
SUNY at Buffalo, USA

ABSTRACT
Collaborative search generally uses previously collected search sessions as a resource to help future users to improve their searching by query modification. The recommendation or automatic extension of the query is generally based on the content of the old sessions, or purely the sequence/order of queries/texts in a session, or a combination. However, users with the same expressed query may need different information. The difference may not be topic related. This chapter proposes to enrich the context of query representation to incorporate non-topical properties of user information needs, which the authors believe will improve the results of collaborative search.

INTRODUCTION: HOW MANY QUESTS ARE THERE?
In support of collaboration among individuals whose work involves exploration of data networks (such as the World Wide Web, or sets of intelligence information, or more detailed databases of scientific information), some researchers are examining a problem that I will call “quest reuse”. The central idea is that a person doing such research is on a quest that has specific goals. Those goals are reflected in the search moves and value judgments made by the investigator. This chapter will research the problem of storing in compact form those moves and judgments so that later investigators may exploit them to speed and increase the effectiveness of their own research.

Anyone who puts down a task, and resumes it some time later makes use of the associating powers of the brain, and of various support systems, to get the human mind back in context. The same individual, even when working with the same set of data, might have several possible contexts, and, indeed might have several contexts simultaneously latent in mind while scanning or browsing. We know that the human mind is superb at this scanning and

DOI: 10.4018/978-1-60566-727-0.ch009
associating activity, and do not envision taking the mind out of the loop. We are looking for ways to increase the power and the ease of that mental work. People often say that there are “innumerable contexts.” We know that the exabytes of data flow are quite numerable -- there are simply too many of them. But the number of contexts is some reasonably finite multiple of the number of people that a system would support. In the whole world this is perhaps thousands of billions. Surely many of them are so similar that sharing and reuse would be really worth aiming for.

In an agency or corporate context, there are perhaps thousands of searchers who might form a pool, and they may each have no more than a few hundred contexts that would be of interest to us. These contexts would have enormous overlap. All the people tracking developments in Iran’s nuclear program have only a few contexts: scientific; political; warning analysis; background analysis, etc. The key idea of “Quest Reuse” (QR) is to store and reuse pools of “quest profiles” that are effectively labeled by context, retrievable by others with closely related contexts, and that contain parameter settings which help a support system (a hypothesis testing tool, a search engine, a report generator, etc.) to refine, disambiguate and prioritize what it seeks and what it finds. To give the familiar trivial example, the word “bank” in the aviation context loads more heavily on “change direction” than it does on “financial institution”.

The chapter will address researches at finding out how to represent and store these profiles, and how to retrieve them by similarity rather than simply by name. That is, QR is valuable if I can know that Larry often works on the same problems as I do, and I tell the machine ‘Please load context “Larry 23”.’ But it is priceless if, by the very actions I take, the system can recognize that I would benefit from the context “Abigail 19”, when I have never heard of Abigail. In a sense this is the familiar theme of “finding experts”. But the experts are to be labeled automatically by analysis of what they look for, what they study hard, and what they mark as “worth keeping” [a crucial part of the model], as well as “what they say they are up to”. We believe this is a worth developing area, so that the many minds working on crucial problems can be working more effectively together, communicating through the perfect memories of the not-very-smart systems that they use.

This same problem, in the specific application to counterterrorism intelligence, is known as “shoebox sharing.” The name is a holdover from the days when an analyst would maintain a shoebox containing 3x5 cards or 5x8 cards summarizing interesting bits of information that he or she found during research. Now that such information is stored on the computer, it could in principle be available to other analysts, but it would be simply a waste of their time unless they can make targeted forays into the material to retrieve that which is most relevant to their own present quests.

**BACKGROUND**

In any model of information retrieval, an information retrieval process is divided into three basic parts: collections of information objects, users, and searching techniques. Developments in information technology, especially the personal computer and the Internet, have brought significant changes to the user and the information collection components. However, the techniques do not change much of the fundamental mechanism that represents and matches user needs and information objects much. With the successful application of information retrieval techniques in web searching since early 90’s, the searching techniques have improved greatly, prompted by the huge business interest of searching market. However, the advances were mainly focused on developing ways of improving existing indexing and ranking techniques. For example, PageRank,