Chapter 4.8
Document Search Practices

Karen L. Corral
Arizona State University, USA

Ryan C. LaBrie
Seattle Pacific University, USA

Robert D. St. Louis
Arizona State University, USA

INTRODUCTION

A large portion of the knowledge of most organizations is contained in electronic documents. For users to get pertinent information from the accumulation of stored documents, they need effective document retrieval systems. Unfortunately, electronic document management has fallen into the same trap that electronic data processing fell into: simply automating what previously was done manually. Paper documents were stored in folders in drawers in file cabinets. Electronic documents are stored in folders in directories on disk drives.

The ability to find a document depends on the logic of the filing system, how familiar the individual is with the filing system, and how familiar the individual is with the problem domain of the item being sought. Some persons (e.g., research librarians) are much better than others at organizing and retrieving documents. Rarely, however, is a manager an expert at either storing or retrieving documents. Unfortunately, many electronic filing systems are set up by managers with little or no training on how to organize a filing system, and few tools, other than the Windows Search command, are available to help managers find documents that have been filed.

The filing systems for libraries and knowledge management systems are more sophisticated than the filing systems of most small offices or individual managers. But even libraries and knowledge management systems predominately rely on keyword searching for retrieval. For example, if one visits the Web site for the Journal of Management Information Systems at http://jmis.bentley.edu/keywords/, one notes that the only option available for searching (other than browsing the entire collection) is a keyword search.

Keyword searching has improved over the years. Knowledge seekers have benefited enormously from the ability to search remotely, the increased speed with which searches are conducted, and the ability of the search mechanism to identify variations of the keywords. Nevertheless, keyword searches have significant limitations. In particular, keyword searches cannot return all relevant documents nor can they filter out irrelevant documents. This article briefly reviews the difficulties associated with keyword searches, especially as the number of documents increases, and proposes a way to overcome those limitations.

BACKGROUND

In his 1990 seminal article on business process engineering, Hammer (1990) argues that organizations should use computers to redesign—not just automate—existing business processes. With document management systems, the opposite has been done. Documents were stored in file cabinets in offices or on shelves in libraries, and electronic document storage systems adopted the same basic principles.

Paper documents such as memos, white papers, reports, and so forth were filed based on the value of some specific field (e.g., project name). To retrieve a document, a user needed to know the value of the field which was used to organize the documents. Because of the shear mass of paper that quickly accumulated in any office, duplication for the purpose of access through multiple fields was not encouraged. In highly organized filing systems, cross-references were filed for important documents, resulting in the capacity to find some documents from two or three different fields. However, this was done infrequently, was quite time-consuming when it was done, and was difficult to maintain.

The logic of the paper filing system usually was determined by a secretary or office assistant, who also was the person primarily responsible for retrieving the documents. This person generally had significant knowledge of the content of the documents, and therefore the system worked quite well for that individual. Unfortunately, the system did not work as well for others.

Today, most individuals organize their computer directories in the same manner in which their file cabinets were organized, or even more poorly because they have had little or no training on filing and tend to store all of their folders in the hard drive root directory. While this may be an acceptable strategy for a small set of documents, it is unacceptable when dealing with a large number of documents.

Figure 1 shows the similarity between paper filing systems and simple computerized filing systems. To find a file in the paper system, an individual needs to know which file cabinet to search, which drawer to select, and which folder contains the sought after file. To find a file in the computerized system, an individual follows a very similar strategy. The individual first selects the disk drive to examine, then searches the directory, sub-directory, sub-sub-directory, and so forth, until the file is located. The only advances made to this point are the amount of physical space saved and the ability to use the "find" command.

At the organizational/library level, document management systems require more structure. Generally, documents are organized by hierarchical levels of categories. For example, with the Dewey Decimal system, documents associated with "technology (applied sciences)" are grouped together. Within that category, "management" is separate from "manufacturing" and so on. A major benefit of this method of organization is that once the individual arrives at one document on the topic of interest, other potentially relevant documents are located in close proximity and are easily browsed for relevance. Everyone who has visited a library has located additional relevant books by browsing the library's physical stacks.

Indexes make cross-referencing of materials possible. Though not physically stored together on