Chapter 7

Strategies for Document Management

Karen Corral
Boise State University, USA

David Schuff
Temple University, USA

Gregory Schymik
Arizona State University, USA

Robert St. Louis
Arizona State University, USA

ABSTRACT

Keyword search has failed to adequately meet the needs of enterprise users. This is largely due to the size of document stores, the distribution of word frequencies, and the indeterminate nature of languages. The authors argue a different approach needs to be taken, and draw on the successes of dimensional data modeling and subject indexing to propose a solution. They test our solution by performing search queries on a large research database. By incorporating readily available subject indexes into the search process, they obtain order of magnitude improvements in the performance of search queries. Their performance measure is the ratio of the number of documents returned without using subject indexes to the number of documents returned when subject indexes are used. The authors explain why the observed tenfold improvement in search performance on our research database can be expected to occur for searches on a wide variety of enterprise document stores.

INTRODUCTION

The amount of information contained within an organization's documents, and the value of that information, is staggering. According to IDC, as of 2007 there were over 281 billion gigabytes of digital information. By 2011, they estimate that this will increase by a factor of 10, and the number of "information assets" will reach 20 quadrillion (Gantz et al., 2008). These information assets are a diverse set of artifacts including everything from television transmissions, to text-based documents, to digital images. BrightPlanet, in a 2005 white paper, estimated that the value of information

DOI: 10.4018/978-1-4666-0279-3.ch007

stored in corporate documents in the United States alone was $3.3 trillion (BrightPlanet, 2005).

Documents and data are important components of these "information assets," but to say that businesses are drowning in a sea of documents is not an exaggeration. In fact, 80% of organizational data is not stored in a traditional database (Olsen, 2003), but instead is stored as a collection of text documents including web pages, word processing documents, electronic mail, and spreadsheets. In this environment, it is extremely difficult to retrieve the right information in a timely manner. One study found that employees spend up to 15% of their time reading information, but up to 50% of their time just finding it (BrightPlanet, 2005). As of today, almost everyone would agree that document management simply has not worked.

The ability to effectively store and retrieve data is well within the reach of most companies. The technologies that underlie databases and data warehouses are well established, the required software and hardware are ubiquitous and affordable, and both managers and their staff are accustomed to interacting with databases and data warehouses. The ability to effectively store and retrieve documents, on the other hand, appears to be beyond the reach of almost all companies. There is little to no agreement on the best way to store documents, and organizations are not content with their ability to locate documents that have been stored. Fortunately, there is reason to be hopeful about the future, and surprisingly that hope does not rest with full-text search engines (such as Google). Instead, managers must make strategic decisions about how their documents are to be stored.

THE CURRENT APPROACH

Apple, Microsoft, and Google all are marketing software that is designed to facilitate document retrieval. Apple claims that its Spotlight search tool "can find anything on your computer as quickly as you type" (Apple, 2009). Moreover, they claim "you always find what you are looking for, even if you don't know where to look." Microsoft and Google make similar claims for their desktop search engines (Google, 2009b; Microsoft, 2009). Google promotes that its search tool "puts your information easily within your reach and frees you from having to manually organize your files, emails and bookmarks" (Google, 2009b).

These claims are quite impressive, and might lead one to believe that the document management problem has been solved. The claims are even more impressive when one stops to consider that the individual or company that owns the documents does not need to add any metadata or structure to the documents before storing them. In fact, for these search engines both organization and format are irrelevant. The artifacts (files, emails, contacts, images, calendars, music, etc.) simply have to be stored on a device accessible by a personal computer or a server. It makes no difference whether the documents are placed in a single folder, or stored in an elaborate hierarchical structure. The presumption is that between the content itself and the artifact's metadata (owner, date created, size, file type, etc.), there is sufficient information to enable retrieval.

Google already has an enterprise version of its Desktop Search tool that provides textual document search across the multiple physical devices on a corporate network (Google, 2009a). Microsoft's Windows Search tool has similar functionality (Microsoft, 2009). Apple's Spotlight search, integrated with its OSX operating system, has the ability to search other networked Macs. The underlying assertion is that everyone will be able to find anything they need whether it is located on the web, on their own PC, or on a corporate server. In other words, if it is digitally stored, you will be able to find it. This sounds too good to be true – and it is!