Chapter IX
A Model for Estimating the Savings from Dimensional vs. Keyword Search

Karen Corral
Boise State University, USA

David Schuff
Temple University, USA

Robert D. St. Louis
Arizona State University, USA

Ozgur Turetken
Ryerson University, Canada

ABSTRACT

Inefficient and ineffective search is widely recognized as a problem for businesses. The shortcomings of keyword searches have been elaborated upon by many authors, and many enhancements to keyword searches have been proposed. To date, however, no one has provided a quantitative model or systematic process for evaluating the savings that accrue from enhanced search procedures. This paper presents a model for estimating the total cost to a company of relying on keyword searches versus a dimensional search approach. The model is based on the Zipf-Mandelbrot law in quantitative linguistics. Our analysis of the model shows that a surprisingly small number of searches are required to justify the cost associated with encoding the metadata necessary to support a dimensional search engine. The results imply that it is cost effective for almost any business organization to implement a dimensional search strategy.

INTRODUCTION

People spend a tremendous amount of time searching for information. One estimate puts the average employee's time at 3-1/2 hours a week for unsuccessful searches (Ultraseek, 2006). For a 1,000 employee company, that works out to $9.7 million a year for just the cost of salary (Ultraseek, 2006). Some estimates put the cost as high as $33 million annually per company when taking into consideration the costs of recreating the information not found (Thompson, 2004). Furthermore, between 60-80% of queries over an intranet (as opposed to the Internet) are for material that the searcher has previously seen (Mukherjee and Mao, 2004).

Keyword search has several well-known problems (for a review, see Blair, 2002), but its advantage over other methods is that once the documents have been saved, there is no additional work that the user has to perform. One alternative to keyword search is dimensional search. Dimensional search eliminates the ambiguity of words (which causes so many of the problems for keyword search) through the use of pre-defined categories (dimensions) to define documents as well as finite sets of possible values for each category. It has been demonstrated that dimensional search reduces the number of irrelevant documents returned in the result set (LaBrie, 2004). However, there is a significant, up-front, time investment that has to be made for dimensional search. In particular, meta-data must be stored about each document, and much of this information must be determined and entered by a human user. So the question becomes, is the increased retrieval accuracy worth the initial cost of categorizing documents?

The content management market was estimated to be over $1 billion in 2003 (Dunwoodie, 2004), and to have grown 9.7% in 2006 (Webster, 2007). Vendors of this software make quite amazing claims about the efficacy of their software, yet for all the money being spent by companies, there has been little academic work done to evaluate these systems. We want to determine the cost, in time, of performing a keyword search versus the cost, in time, of performing a dimensional search, including the initial time-investment. Factors that affect the overall cost of searching include the start-up costs of any content management system, the size of the library (it is much easier to exhaustively search a small library than a large library), the size of the documents in the library (books are more difficult to search than are e-mail messages), and the cost of not finding the document.

While evaluating the best approach to studying this question, we considered a number of research methodologies. A case study approach to this problem, which is largely what IDC, Gartner and other commercial information providers use, would be hampered by a lack of generalizability. Also, attempting to collect data on an employee's search could be considered invasive by the employee. If employees know that their time and actions are being tracked, they might elect to perform searches outside of such data collection, out of concern that the collected data might be used to evaluate their work rather than the content management software. Moreover, drawing data from a survey of content management product users makes comparison of such data difficult as the nature of searches might vary considerably by company as well as by user. And there is the additional concern that users might not have an accurate sense of the time or the effectiveness of their searches.

An experiment would need to consider all the above factors, plus ensure the proper motivation of the users. For these reasons, we elected to use an analytical modeling approach, which allows us to use different values for variables and examine their impact on search cost. From our model we were able to determine the break-even point, in terms of the number of searches, at which dimensional search becomes more cost effective than keyword search. That is, we were able to