Chapter II

Search Engines, Relevancy, and the World Wide Web

Wendy Lucas
Bentley College, USA

This chapter presents a survey and discussion of the relevancy rankings assigned by search engines to pages in their indices. An examination of methodologies whose goal it is to improve the relevancy of search results follows. The chapter concludes with a look at emerging trends in search engine technologies and directions for future research.

INTRODUCTION

Recent estimates for the size of the World Wide Web range from 275 million distinct, static pages, growing by about 20 million pages per month (Bharat & Border, 1999), to 800 million publicly “indexable” pages (Lawrence & Giles, 1999), where indexable refers to those pages that are available to search engines for indexing. While estimates vary, it is undeniable that the exponential growth rate of the Web greatly complicates the task of finding relevant pages. Organizing and managing these pages therefore assumes critical importance. This role has largely been borne by search engines, which impose structure on the Web by indexing its contents and providing directed
access via search interfaces.

An estimated 85% of users turn to search engines for finding information on the Web (Graphic Visualization and Usability Center, 1998). As anyone who has used a search engine knows, submitting a query to an engine often results in the listing of thousands of links to “matching pages” of dubious quality. Links to retrieved pages are listed in decreasing order of relevance, with relevancy ratings determined by proprietary algorithms. Typically, links to 10 Web pages are displayed on each page of search results. Pages whose links appear on the first page of results are the most likely to be viewed (Silverstein, Henzinger, Marais, & Moricz, 1998; Spink, Chang, & Goz, 1999). Yet, there is no guarantee that the most relevant pages will be those appearing in the “top ten” (Gordon & Pathak, 1999; Gudivada, Raghavan, Grosky, & Kasanagottu, 1997). A high-quality page that truly matches the user’s information needs may therefore never be discovered if the ranking algorithm used by the search engine erroneously places its link beyond the first page or two of results with other, less relevant links appearing before it.

The focus of this chapter is on the relevancy of search results returned by the most popular search engines in response to users’ queries. Descriptions of the functions performed by search engines and a review of current literature analyzing their performance are given in the next section. A preliminary study that evaluates methodologies for improving relevance, which was first presented in (Lucas, 2000), follows. This study investigates the relevance of search engine results to user-entered queries and then examines a method for reducing the incidence of irrelevant matches. Approaches that rely on the internal and external structure of Web documents are supported by recent research. Emerging trends in search engine technologies, their impact on text databases, and directions for future research are also explored.

BACKGROUND

Many of today’s search engines offer two ways for users to find information on the Web. One is by submitting a query whose content is then matched to the contents of an electronically built and maintained index. The other is by making use of directories that are created by human volunteers for classifying Web sites. As the Web continues to expand exponentially, efficient means for classifying and managing the textual contents of this largest of databases will necessitate the continued use of electronic measures. These measures are therefore the focus of this chapter.
Goal Modeling in Requirements Engineering: Analysis and Critique of Current Methods
[www.igi-global.com/chapter/goal-modeling-requirements-engineering/23011?camid=4v1a](www.igi-global.com/chapter/goal-modeling-requirements-engineering/23011?camid=4v1a)

Conflicts, Compromises and Political Decisions: Methodological Challenges of Enterprise-Wide E-Business Architecture
[www.igi-global.com/article/conflicts-compromises-political-decisions/3380?camid=4v1a](www.igi-global.com/article/conflicts-compromises-political-decisions/3380?camid=4v1a)