Information Retrieval on the Web

MEI KOBAYASHI and KOICHI TAKEDA

IBM Research

In this paper we review studies of the growth of the Internet and technologies that are useful for information search and retrieval on the Web. We present data on the Internet from several different sources, e.g., current as well as projected number of users, hosts, and Web sites. Although numerical figures vary, overall trends cited by the sources are consistent and point to exponential growth in the past and in the coming decade. Hence it is not surprising that about 85% of Internet users surveyed claim using search engines and search services to find specific information. The same surveys show, however, that users are not satisfied with the performance of the current generation of search engines; the slow retrieval speed, communication delays, and poor quality of retrieved results (e.g., noise and broken links) are commonly cited problems. We discuss the development of new techniques targeted to resolve some of the problems associated with Web-based information retrieval,and speculate on future trends.

Categories and Subject Descriptors: G.1.3 [**Numerical Analysis**]: Numerical Linear Algebra—*Eigenvalues and eigenvectors* (direct and iterative methods); *Singular value decomposition*; *Sparse, structured and very large systems* (direct and iterative methods); G.1.1 [**Numerical Analysis**]: Interpolation; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*; *Retrieval models*; *Search process*; H.m [**Information Systems**]: Miscellaneous

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Clustering, indexing, information retrieval, Internet, knowledge management, search engine, World Wide Web

1. INTRODUCTION

We review some notable studies on the growth of the Internet and on technologies useful for information search and retrieval on the Web. Writing about the Web is a challenging task for several reasons, of which we mention three. First, its dynamic nature guarantees that at least some portions of any manuscript on the subject will be out-of-date before it reaches the intended audience, particularly URLs that are referenced. Second, a comprehensive coverage of all of the important topics is impossible, because so many new ideas are constantly being proposed and are either quickly accepted into the Internet mainstream or rejected. Finally, as with any review paper, there is a strong bias

Authors' address: Tokyo Research Laboratory, IBM Research, 1623–14 Shimotsuruma, Yamato-shi, Kanagawa-ken, 242–8502, Japan; email: mei_kobayashi@jp.ibm.com; kohichi_takeda@jp.ibm.com.

CONTENTS

in presenting topics closely related to the authors' background, and giving only cursory treatment to those of which they are relatively ignorant. In an attempt to compensate for oversights and biases, references to relevant works that describe or review concepts in depth will be given whenever possible. This being said, we begin with references to several excellent books that cover a variety of topics in information management and retrieval. They include *Information Retrieval and Hypertext* [Agosti and Smeaton 1996]; *Modern Information Retrieval* [Baeza-Yates and Ribeiro-Neto 1999]; *Text Retrieval and Filtering: Analytic Models of Performance* [Losee 1998]; *Natural Language Information Retrieval* [Strzalkowski 1999]; and *Managing Gigabytes* [Witten et al. 1994]. Some older, classic texts, which are slightly outdated, include *Information Retrieval* [Frakes and Baeza-Yates 1992]; *Information Storage and Retrieval* [Korfhage 1997]; *Intelligent Multmedia Information Retrieval* [Maybury 1997]; *Introduction to Modern Information Retrieval* [Salton and McGill 1983]; and *Readings in Information Retrieval* [Jones and Willett 1977].

Additional references are to special journal issues on search engines on the Internet [Scientific American 1997]; digital libraries [CACM 1998]; digital libraries, representation and retrieval [IEEE 1996b]; the next generation graphical user interfaces (GUIs) [CACM 1994]; Internet technologies [CACM 1994; IEEE 1999]; and knowledge discovery [CACM 1999]. Some notable survey papers are those by Chakrabarti and Rajagopalan [1997]; Faloutsos and Oard [1995]; Feldman [1998]; Gudivada et al. [1997]; Leighton and Srivastava [1997]; Lawrence and Giles [1998b; 1999b]; and Raghavan [1997]. Extensive, up-to-date coverage of topics in Web-based information retrieval and knowledge management can be found in the proceedings of several conferences, such as: the *International World Wide Web Conferences* [WWW Conferences 2000] and the Association for Computing Machinery's Special Interest Group on Computer-Human Interaction [ACM SIGCHI] and Special Interest Group on Information Retrieval [ACM SIGIR] conferences <acm.org>. A list of papers and Web pages that review and compare Web search tools are maintained at several sites, including Boutell's World Wide Web FAQ <boutell.com/faq/>; Hamline University's <web.hamline.edu/administration/libraries/search/comparisons.html>; Kuhn's pages (in German) <gwdg.de/hkuhn1/pagesuch.html#vl>; Maire's pages (in French) <imaginet.fr/ime/search.htm>; Princeton University's <cs.princeton.edu/html/search.html>; U.C. Berkeley's <sunsite.berkeley.edu/help/searchdetails.html>; and Yahoo!'s pages on search engines <yahoo.com/computers and internet/internet/world wide web>. The historical development of information retrieval is documented in a number of sources: Baeza-Yates and Ribeiro-Neto [1999]; Cleverdon [1970]; Faloutsos and Oard [1995]; Salton [1970]; and van Rijsbergen [1979]. Historical accounts of the Web and Web search technologies are given in Berners-Lee et al. [1994] and Schatz [1997].

This paper is organized as follows. In the remainder of this section, we discuss and point to references on ratings of search engines and their features, the growth of information available on the Internet, and the growth in users. In the second section we present tools for Web-based information retrieval. These