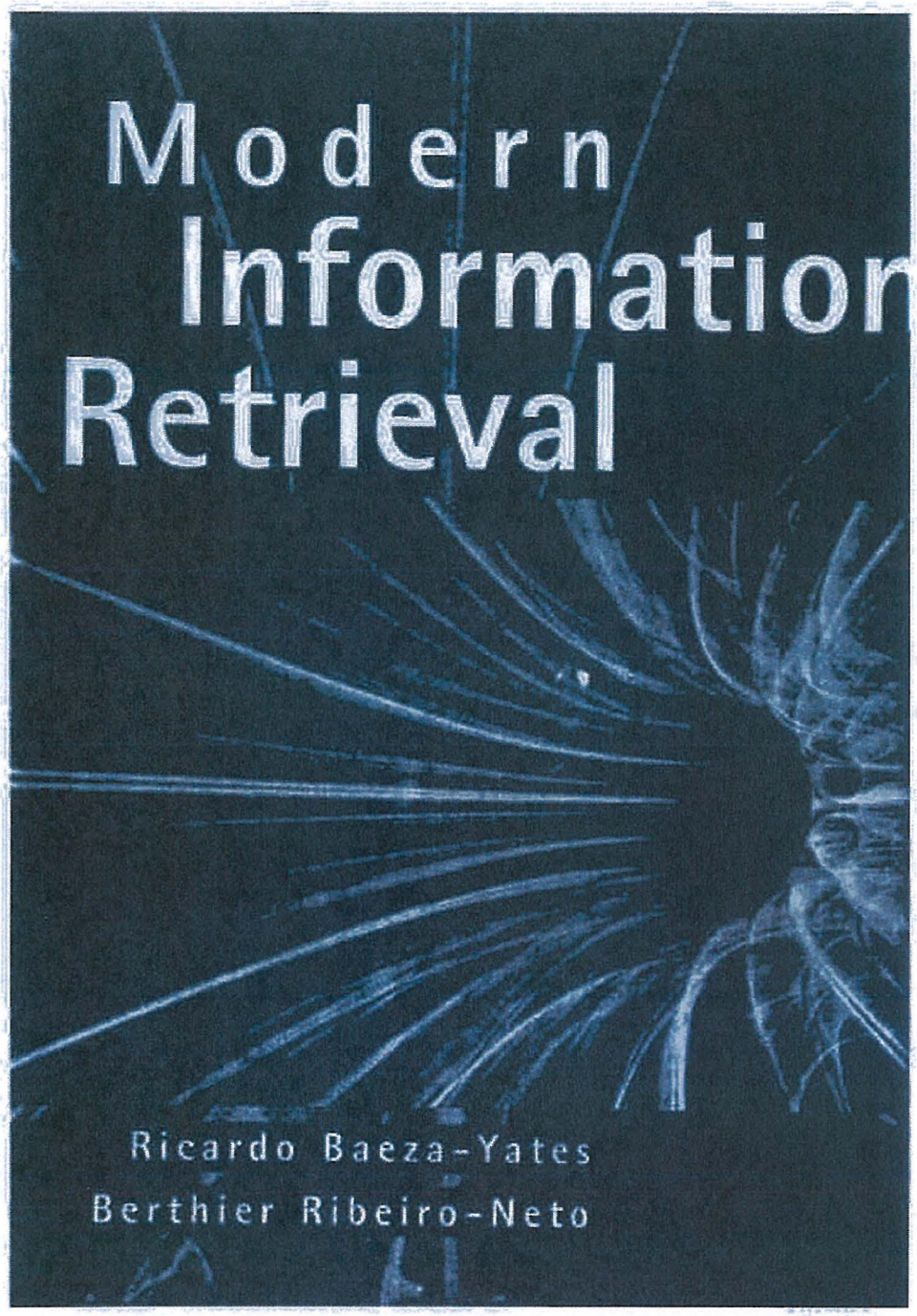


Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the World.
Visit us on the World Wide Web at
http://www.pearsoneduc.com

Typeset in Computer Modern by 56
Printed and bound in the United States of America

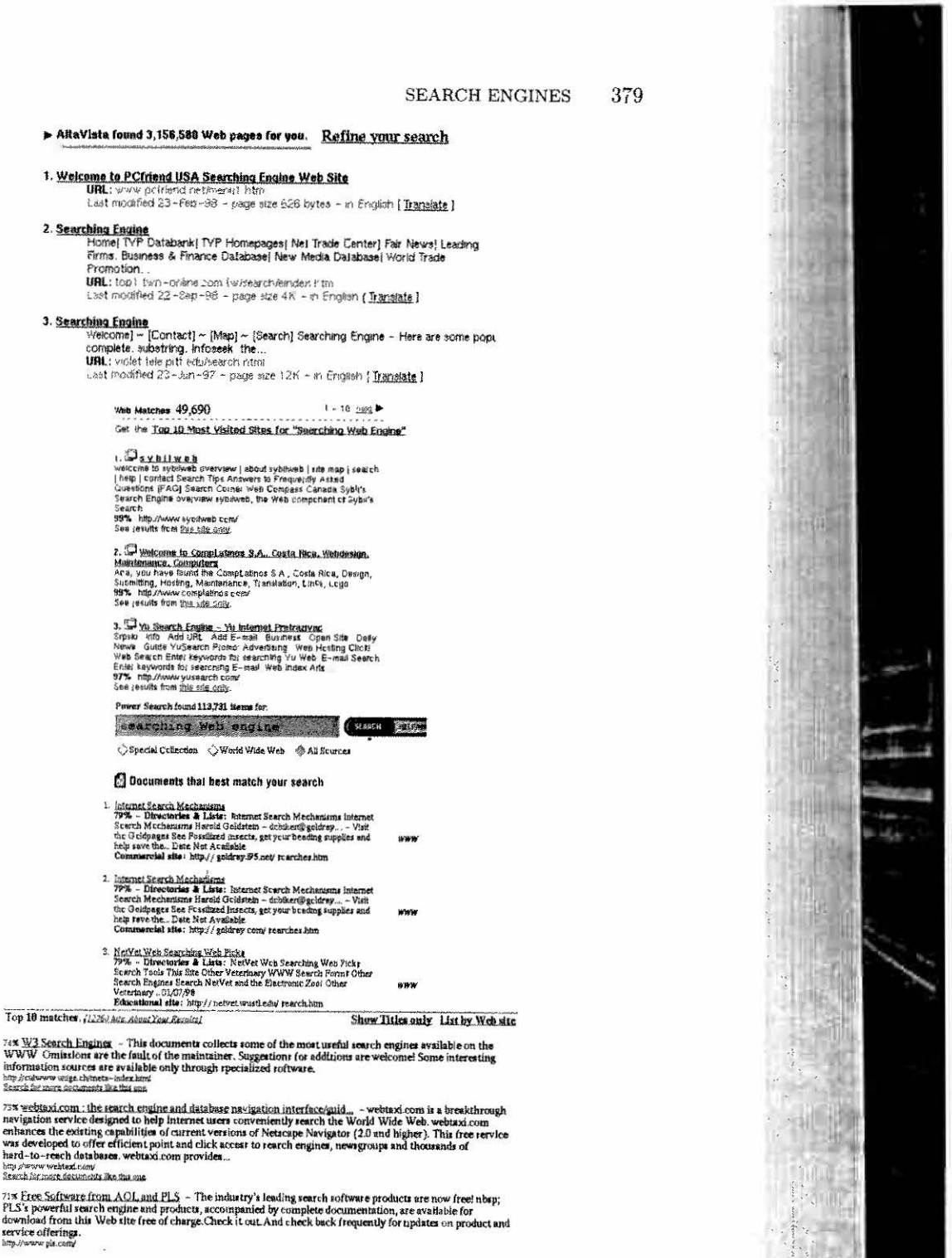
First printed 1999

ISBN 0-201-39829-X

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloguing-in-Publication Data
Baeza-Yates, R.(Ricardo)
Modern information retrieval / Ricardo Baeza-Yates, Berthier Ribeiro-Neto.
p. cm.
Includes bibliographical references and index.
ISBN 0-201-39829-X
1. Information storage and retieval systems. I. Ribeiro, Berthier de Araújo Neto, 1960- . II.Title.
Z667.B34 1999
025.04–dc21 99-10033
CIP

10 9 8 7 6 5 4 3
04 03 02 01 00



▶ **AltaVista found 3,156,580 Web pages for you.** Refine your search

1. **Welcome to PCfriend USA Searching Engine Web Site**
   URL: www.pcfriend.net/search.htm
   Last modified 23-Feb-98 - page size 626 bytes - in English [Translate]

2. **Searching Engine**
   Home| TVP Databank| TVP Homepages| Net Trade Center| Fair News| Leading Firms. Business & Finance Database| New Media Database| World Trade Promotion. . .
   URL: top1.twn-online.com.tw/search/engine.htm
   Last modified 22-Sep-98 - page size 4K - in English [Translate]

3. **Searching Engine**
   Welcome] ~ [Contact] ~ [Map] ~ [Search] Searching Engine - Here are some popu complete. substring. Infoseek the...
   URL: violet.lele.pitt.edu/search.html
   Last modified 23-Jun-97 - page size 12K - in English [Translate]

Web Matches 49,690 1 - 10 next ▶

Get the Top 10 Most Visited Sites for "Searching Web Engine"

1. sybilweb
welcome to sybilweb overview | about sybilweb | site map | search | help | contact Search Tips Answers to Frequently Asked Questions (FAQ) Search Comsi Web Compass Canada Sybil's Search Engine overview sybilweb, the Web component of Sybil's Search
99% http://www.sybilweb.com/
See results from this site only.

2. Welcome to CompuLatinos S.A., Costa Rica, Webdesign, Maintenance, Computers
Arca, you have found the CompuLatinos S.A., Costa Rica, Design, Submitting, Hosting, Maintenance, Translation, Links, Logo
99% http://www.compulatinos.com/
See results from this site only.

3. Yu Search Engine - Yu Internet Pretraživač
Srpski Info Add URL Add E-mail Business Open Site Daily News Guide YuSearch Pricing Advertising Web Hosting Click! Web Search Enter keywords for searching Yu Web E-mail Search Enter keywords for searching E-mail Web Index Arts
97% http://www.yusearch.com/
See results from this site only.

Power Search found 113,731 items for:

searching Web engine SEARCH CLEAR

Special Collection World Wide Web All Sources

Documents that best match your search

1. Internet Search Mechanisms
79% - **Directories & Lists:** Internet Search Mechanisms Internet Search Mechanisms Harold Goldstein - dchbker@goldray... - Visit the Goldpages See Fossilized insects, get your beading supplies and help save the... Date Not Available WWW
Commercial site: http://goldray.95.net/searches.htm

2. Internet Search Mechanisms
79% - **Directories & Lists:** Internet Search Mechanisms Internet Search Mechanisms Harold Goldstein - dchbker@goldray... - Visit the Goldpages See Fossilized Insects, get your beading supplies and help save the... Date Not Available WWW
Commercial site: http://goldray.com/searches.htm

3. NetVet Web Searching Web Picks
79% - **Directories & Lists:** NetVet Web Searching Web Picks Search Tools This Site Other Veterinary WWW Search Forms Other Search Engines Search NetVet and the Electronic Zoo! Other Veterinary .. 01/07/98 WWW
Educational site: http://netvet.wustl.edu/search.htm

Top 10 matches. (1126 hits About Your Results) Show Titles only List by Web site

74% W3 Search Engines - This documents collects some of the most useful search engines available on the WWW Omissions are the fault of the maintainer. Suggestions (or additions are welcome! Some interesting information sources are available only through specialized software.
http://cuiwww.unige.ch/meta-index.html
Search for more documents like this one

73% webtaxi.com : the search engine and database navigation interface/guid... - webtaxi.com is a breakthrough navigation service designed to help Internet users conveniently search the World Wide Web. webtaxi.com enhances the existing capabilities of current versions of Netscape Navigator (2.0 and higher). This free service was developed to offer efficient point and click access to search engines, newsgroups and thousands of hard-to-reach databases. webtaxi.com provides...
http://www.webtaxi.com/
Search for more documents like this one

71% Free Software from AOL and PLS - The industry's leading search software products are now free! nbsp; PLS's powerful search engine and products, accompanied by complete documentation, are available for download from this Web site free of charge.Check it out.And check back frequently for updates on product and service offerings.
http://www.pls.com/

Figure 13.7 Output for the query `searching and Web and engine` for the four main search engines; from top to bottom: AltaVista, HotBot, NorthernLight, and Excite.

The user can also refine the query by constructing more complex queries based on the previous answer.

The Web pages retrieved by the search engine in response to a user query are ranked, usually using statistics related to the terms in the query. In some cases this may not have any meaning, because relevance is not fully correlated with statistics about term occurrence within the collection. Some search engines also taking into account terms included in metatags or the title, or the popularity of a Web page to improve the ranking. This topic is covered next.

13.4.4 Ranking

Most search engines use variations of the Boolean or vector model (see Chapter 2) to do ranking. As with searching, ranking has to be performed without accessing the text, just the index. There is not much public information about the specific ranking algorithms used by current search engines. Further, it is difficult to compare fairly different search engines given their differences, and continuous improvements. More important, it is almost impossible to measure recall, as the number of relevant pages can be quite large for simple queries. Some inconclusive studies include [327, 498].

Ywono and Lee [844] propose three ranking algorithms in addition to the classical tf-idf scheme (see Chapter 2). They are called Boolean spread, vector spread, and most-cited. The first two are the normal ranking algorithms of the Boolean and vector model extended to include pages pointed to by a page in the answer or pages that point to a page in the answer. The third, most-cited, is based only on the terms included in pages having a link to the pages in the answer. A comparison of these techniques considering 56 queries over a collection of 2400 Web pages indicates that the vector model yields a better recall-precision curve, with an average precision of 75%.

Some of the new ranking algorithms also use hyperlink information. This is an important difference between the Web and normal IR databases. The number of hyperlinks that point to a page provides a measure of its popularity and quality. Also, many links in common between pages or pages referenced by the same page often indicates a relationship between those pages. We now present three examples of ranking techniques that exploit these facts, but they differ in that two of them depend on the query and the last does not.

The first is WebQuery [148], which also allows visual browsing of Web pages. WebQuery takes a set of Web pages (for example, the answer to a query) and ranks them based on how connected each Web page is. Additionally, it extends the set by finding Web pages that are highly connected to the original set. A related approach is presented by Li [512].

A better idea is due to Kleinberg [444] and used in HITS (Hypertext Induced Topic Search). This ranking scheme depends on the query and considers the set of pages S that point to or are pointed by pages in the answer. Pages that have many links pointing to them in S are called authorities (that is, they should have relevant content). Pages that have many outgoing links are called hubs (they should point to similar content). A positive two-way feedback exists:

better authority pages come from incoming edges from good hubs and better hub pages come from outgoing edges to good authorities. Let $H(p)$ and $A(p)$ be the hub and authority value of page p. These values are defined such that the following equations are satisfied for all pages p:

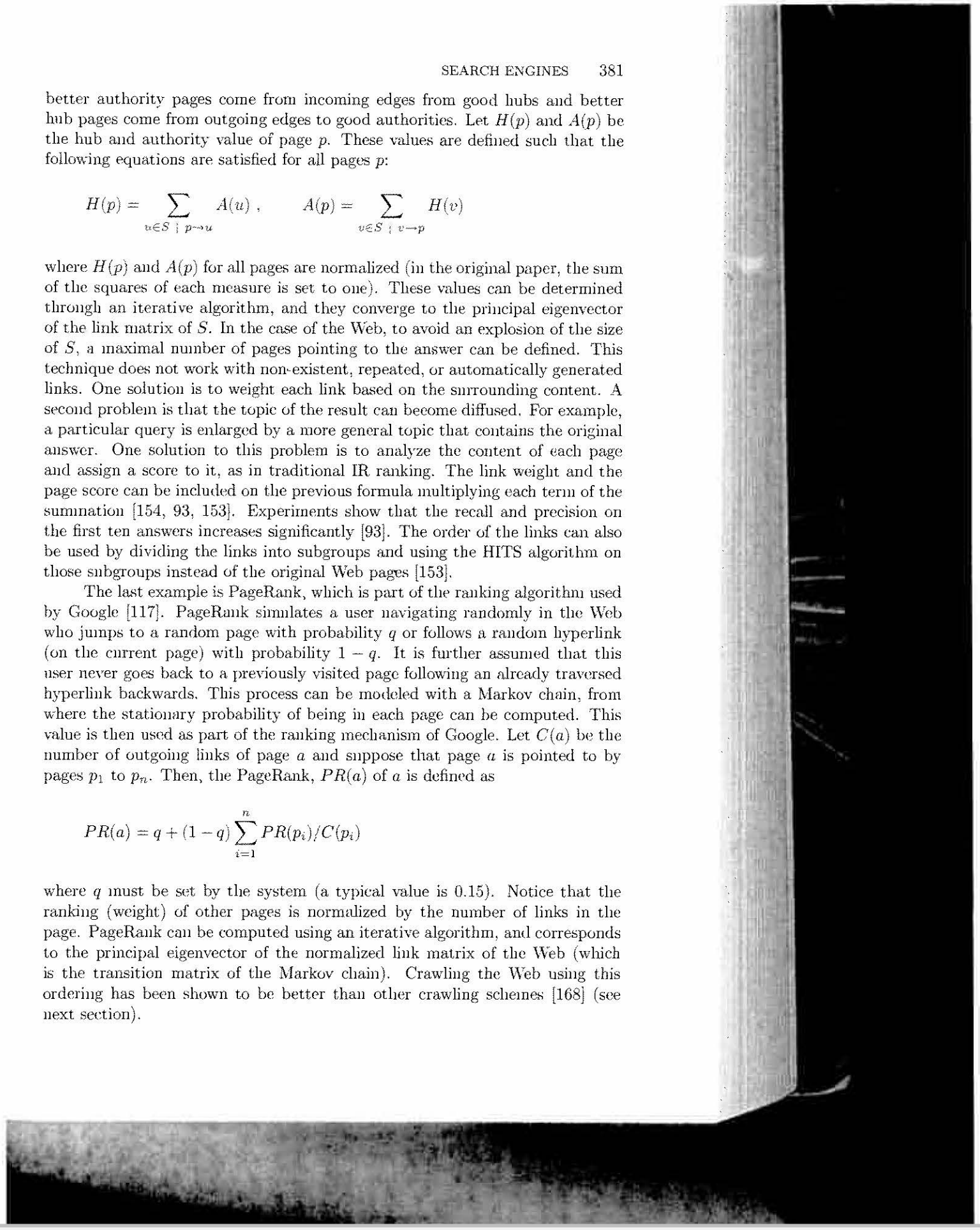
$$H(p) = \sum_{u \in S \mid p \to u} A(u) , \qquad A(p) = \sum_{v \in S \mid v \to p} H(v)$$

where $H(p)$ and $A(p)$ for all pages are normalized (in the original paper, the sum of the squares of each measure is set to one). These values can be determined through an iterative algorithm, and they converge to the principal eigenvector of the link matrix of S. In the case of the Web, to avoid an explosion of the size of S, a maximal number of pages pointing to the answer can be defined. This technique does not work with non-existent, repeated, or automatically generated links. One solution is to weight each link based on the surrounding content. A second problem is that the topic of the result can become diffused. For example, a particular query is enlarged by a more general topic that contains the original answer. One solution to this problem is to analyze the content of each page and assign a score to it, as in traditional IR ranking. The link weight and the page score can be included on the previous formula multiplying each term of the summation [154, 93, 153]. Experiments show that the recall and precision on the first ten answers increases significantly [93]. The order of the links can also be used by dividing the links into subgroups and using the HITS algorithm on those subgroups instead of the original Web pages [153].

The last example is PageRank, which is part of the ranking algorithm used by Google [117]. PageRank simulates a user navigating randomly in the Web who jumps to a random page with probability q or follows a random hyperlink (on the current page) with probability $1 - q$. It is further assumed that this user never goes back to a previously visited page following an already traversed hyperlink backwards. This process can be modeled with a Markov chain, from where the stationary probability of being in each page can be computed. This value is then used as part of the ranking mechanism of Google. Let $C(a)$ be the number of outgoing links of page a and suppose that page a is pointed to by pages p_1 to p_n. Then, the PageRank, $PR(a)$ of a is defined as

$$PR(a) = q + (1 - q) \sum_{i=1}^{n} PR(p_i)/C(p_i)$$

where q must be set by the system (a typical value is 0.15). Notice that the ranking (weight) of other pages is normalized by the number of links in the page. PageRank can be computed using an iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the Web (which is the transition matrix of the Markov chain). Crawling the Web using this ordering has been shown to be better than other crawling schemes [168] (see next section).



DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.