Advanced Metasearch Engine Technology

© Springer Nature Switzerland AG 2022
Reprint of original edition © Morgan & Claypool, 2011

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Advanced Metasearch Engine Technology

Weiyi Meng and Clement T. Yu

ISBN: 978-3-031-00715-6 paperback
ISBN: 978-3-031-01843-5 ebook

DOI 10.1007/978-3-031-01843-5

A Publication in the Springer series
SYNTHESIS LECTURES ON DATA MANAGEMENT

Lecture #11
Series Editor: M. Tamer Özsu, *University of Waterloo*
Series ISSN
Synthesis Lectures on Data Management
Print 2153-5418 Electronic 2153-5426

Synthesis Lectures on Data Management

Editor

M. Tamer Özsu, ***University of Waterloo***

Synthesis Lectures on Data Management is edited by Tamer Özsu of the University of Waterloo. The series will publish 50- to 125 page publications on topics pertaining to data management. The scope will largely follow the purview of premier information and computer science conferences, such as ACM SIGMOD, VLDB, ICDE, PODS, ICDT, and ACM KDD. Potential topics include, but not are limited to: query languages, database system architectures, transaction management, data warehousing, XML and databases, data stream systems, wide scale data distribution, multimedia data management, data mining, and related subjects.

Advanced Metasearch Engine Technology
Weiyi Meng and Clement T. Yu
2010

Web Page Recommendation Models: Theory and Algorithms
Sule Gündüz-Ögüdücü
2010

Multidimensional Databases and Data Warehousing
Christian S. Jensen, Torben Bach Pedersen, and Christian Thomsen
2010

Database Replication
Bettina Kemme, Ricardo Jimenez Peris, and Marta Patino-Martinez
2010

Relational and XML Data Exchange
Marcelo Arenas, Pablo Barcelo, Leonid Libkin, and Filip Murlak
2010

User-Centered Data Management
Tiziana Catarci, Alan Dix, Stephen Kimani, and Giuseppe Santucci
2010

Data Stream Management
Lukasz Golab and M. Tamer Özsu
2010

Access Control in Data Management Systems
Elena Ferrari
2010

An Introduction to Duplicate Detection
Felix Naumann and Melanie Herschel
2010

Privacy-Preserving Data Publishing: An Overview
Raymond Chi-Wing Wong and Ada Wai-Chee Fu
2010

Keyword Search in Databases
Jeffrey Xu Yu, Lu Qin, and Lijun Chang
2009

Advanced Metasearch Engine Technology

Weiyi Meng
SUNY Binghamtom

Clement T. Yu
University of Illinois at Chicago

SYNTHESIS LECTURES ON DATA MANAGEMENT #11

ABSTRACT

Among the search tools currently on the Web, search engines are the most well known thanks to the popularity of major search engines such as Google and Yahoo!. While extremely successful, these major search engines do have serious limitations.

This book introduces large-scale metasearch engine technology, which has the potential to overcome the limitations of the major search engines. Essentially, a metasearch engine is a search system that supports unified access to multiple existing search engines by passing the queries it receives to its component search engines and aggregating the returned results into a single ranked list. A large-scale metasearch engine has thousands or more component search engines. While metasearch engines were initially motivated by their ability to combine the search coverage of multiple search engines, there are also other benefits such as the potential to obtain better and fresher results and to reach the Deep Web.

The following major components of large-scale metasearch engines will be discussed in detail in this book: search engine selection, search engine incorporation, and result merging. Highly scalable and automated solutions for these components are emphasized. The authors make a strong case for the viability of the large-scale metasearch engine technology as a competitive technology for Web search.

KEYWORDS

metasearch engine, large-scale metasearch engine, search broker, distributed information retrieval, federated search system, search engine selection, database selection, search result extraction, wrapper generation, result merging, collection fusion

Contents

Preface ix

Acknowledgments xi

1 Introduction 1

1.1 Finding Information on the Web 2
1.1.1 Browsing 2
1.1.2 Searching 3
1.2 A Brief Overview of Text Retrieval 5
1.2.1 System Architecture 5
1.2.2 Document Representation 6
1.2.3 Document-Query Matching 7
1.2.4 Query Evaluation 8
1.2.5 Retrieval Effectiveness Measures 10
1.3 A Brief Overview of Search Engine Technology 11
1.3.1 Special Characteristics of the Web 11
1.3.2 Web Crawler 12
1.3.3 Utilizing Tag Information 13
1.3.4 Utilizing Link Information 14
1.3.5 Result Organization 18
1.4 Book Overview 19

2 Metasearch Engine Architecture 21

2.1 System Architecture 21
2.2 Why Metasearch Engine Technology? 26
2.3 Challenging Environment 31
2.3.1 Heterogeneities and Their Impact 31
2.3.2 Standardization Efforts 34

3 Search Engine Selection 37

3.1 Rough Representative Approaches 38
3.2 Learning-Based Approaches 39

3.3 Sample Document-Based Approaches 42
3.4 Statistical Representative Approaches 46
3.4.1 D-WISE 46
3.4.2 CORI Net 47
3.4.3 gGlOSS 49
3.4.4 Number of Potentially Useful Documents 51
3.4.5 Similarity of the Most Similar Document 53
3.4.6 Search Engine Representative Generation 56

4 Search Engine Incorporation 61
4.1 Search Engine Connection 61
4.1.1 HTML Form Tag for Search Engines 61
4.1.2 Automatic Search Engine Connection 64
4.2 Search Result Extraction 68
4.2.1 Semiautomatic Wrapper Generation 70
4.2.2 Automatic Wrapper Generation 73

5 Result Merging 85
5.1 Merging Based on Full Document Content 87
5.2 Merging Based on Search Result Records 89
5.3 Merging Based on Local Ranks of Results 93
5.3.1 Round-Robin Based Methods 94
5.3.2 Similarity Conversion Based Methods 95
5.3.3 Voting Based Methods 97
5.3.4 Machine Learning Based Method 101

6 Summary and Future Research 103

Bibliography 107

Authors' Biographies 117

Preface

The World Wide Web has become the largest information source in recent years and developing advanced search tools has been a key research and development activity of the Internet technology. Among the search tools currently on the Web, search engines are the most well known thanks to the popularity of major search engines such as Google and Yahoo!. While extremely successful, these major search engines do have serious limitations. For example, each of them only covers a small fraction of the full contents available on the Web, and their crawling based technology has difficulty in fully reaching the so-called *deep Web*, although much progress on this aspect has been made recently, and in keeping up with the changes and expansion of the Web content.

In this book, we introduce the metasearch engine technology, which has the potential to overcome the limitations of the major search engines. A metasearch engine is a search system that supports unified access to some existing search engines. Essentially, a metasearch engine passes the queries it receives to other search engines, and when the results are returned from the invoked search engines, it aggregates the results into a single ranked list and presents them to the user. While metasearch engines were initially motivated by their ability to combine the search coverage of multiple search engines, there are additional benefits such as the potential to obtain better and fresher results and to reach the deep Web.

In this book, we emphasize the concept of *large-scale metasearch engine*. Such a metasearch engine is connected to thousands or more search engines. To build and maintain large-scale metasearch engines, *advanced metasearch search engine technology* with highly scalable and automated solutions for several key components is needed. This book aims to provide an in-depth and comprehensive coverage of advanced metasearch engine technology. We will make a strong case for the viability of the large-scale metasearch engines as a competitive technology for Web search. The following major components of large-scale metasearch engine will be discussed in detail in this book: search engine selection – a component that identifies search engines that are most likely to provide useful results for any give query; search engine incorporation – a component that interacts with various search engines, including passing queries from the metasearch engine to local search engines and extracting the search results from response pages returned from different search engines; and result merging – a component that merges results returned from different search engines into a single ranked list. Advanced metasearch engine technology includes highly accurate and scalable search engine selection algorithms, highly automated search engine incorporation techniques, and highly effective result merging methods.

This book can be used as part of Web technology related courses such as Web data management and information retrieval. It can also be a reference book for IT professionals, especially researchers and developers in the Web search area.

Weiyi Meng and Clement T. Yu
November 2010

Acknowledgments

We like to express our deep gratitude to the series editor of the Synthesis Lectures on Data Management, Dr. M. Tamer Özsu, who read the entire manuscript carefully and provided very valuable and constructive suggestions, which have helped improve the book significantly. We also like to thank Hongkun Zhao and Can Lin for reading parts of the manuscript and providing valuable comments. We are also grateful to the editor of the book, Diane Cerra, for her help in the course of preparing the book.

Weiyi Meng and Clement T. Yu
November 2010