

Relevance Ranking for Vertical Search Engines

Bo Long & Yi Chang *Editors*

MK
MORGAN KAUFMANN

Relevance Ranking for Vertical Search Engines

FIRST EDITION

Bo Long and Yi Chang

LinkedIn Inc., Mountain View, CA, USA

Yahoo! Labs, Sunnyvale, CA, USA



AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Morgan Kaufmann is an imprint of Elsevier

Table of Contents

Cover image

Title page

Copyright

List of Tables

List of figures

About the Editors

List of Contributors

Foreword

1: Introduction

1.1 Defining the Area

1.2 The Content and Organization of This Book

1.3 The Audience for This Book

1.4 Further Reading

2: News Search Ranking

2.1 The Learning-to-Rank Approach

2.2 Joint Learning Approach from Clickthroughs

2.3 News Clustering

2.4 Summary

3: Medical Domain Search Ranking

Introduction

3.1 Search Engines for Electronic Health Records

3.2 Search Behavior Analysis

3.3 Relevance Ranking

3.4 Collaborative Search

3.5 Conclusion

4: Visual Search Ranking

Introduction

4.1 Generic Visual Search System

4.2 Text-Based Search Ranking

4.3 Query Example-Based Search Ranking

4.4 Concept-Based Search Ranking

4.5 Visual Search Reranking

4.6 Learning and Search Ranking

4.7 Conclusions and Future Challenges

5: Mobile Search Ranking

Introduction

5.1 Ranking Signals

5.2 Ranking Heuristics

5.3 Summary and Future Directions

6: Entity Ranking

6.1 An Overview of Entity Ranking

6.2 Background Knowledge

6.3 Feature Space Analysis

6.4 Machine-Learned Ranking for Entities

6.5 Experiments

6.6 Conclusions

7: Multi-Aspect Relevance Ranking

[Introduction](#)

[7.1 Related Work](#)

[7.2 Problem Formulation](#)

[7.3 Learning Aggregation Functions](#)

[7.4 Experiments](#)

[7.5 Conclusions and Future Work](#)

[8: Aggregated Vertical Search](#)

[Introduction](#)

[8.1 Sources of Evidence](#)

[8.2 Combination of Evidence](#)

[8.3 Evaluation](#)

[8.4 Special Topics](#)

[8.5 Conclusion](#)

[9: Cross-Vertical Search Ranking](#)

[Introduction](#)

[9.1 The PCDF Model](#)

[9.2 Algorithm Derivation](#)

[9.3 Experimental Evaluation](#)

[9.4 Related Work](#)

[9.5 Conclusions](#)

[References](#)


[Author Index](#)

[Subject Index](#)

Copyright

Acquiring Editor: Steve Elliot
Editorial Project Manager: Lindsay Lawrence
Project Manager: Punithavathy Govindaradjane
Designer: Maria Inês Cruz
Morgan Kaufmann is an imprint of Elsevier
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2014 Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions .

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

Relevance ranking for vertical search engines / Bo Long, Yi Chang (Editors).

pages cm

Includes bibliographical references and index.

ISBN 978-0-12-407171-1

1. Text processing (Computer science) 2. Sorting (Electronic computers) 3. Relevance. 4. Database searching. 5. Search engines—Programming. I. Long, Bo, editor of compilation.

II. Chang, Yi (Computer expert)

QA76.9.T48R455 2014

025.04—dc23

2013039777

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN: 978-0-12-407171-1


Printed and bound in the United States of America

14 15 16 17 18 10 9 8 7 6 5 4 3 2 1



Working together
to grow libraries in
developing countries

www.elsevier.com • www.bookaid.org

For information on all MK publications visit our website at www.mkp.com 

List of Tables

Number	Table	Page
2.1	Temporal features for URL freshness and query model.	16
2.2	Evaluation corpus.	19
2.3	Feature weights learned by JRFL.	21
2.4	Performance of individual relevance and freshness estimations.	22
2.5	Query intention analysis by the inferred query weight.	23
2.6	Query length distribution under different query categories.	23
2.7	Comparison of random bucket clicks.	24
2.8	Comparison of normal clicks.	24
2.9	Comparison of editorial annotations.	25
2.10	Case study: Degenerated ranking results by JRFL for query “afghanistan.”	25
2.11	Performance of offline clustering system.	33
2.12	Results with various simple offline clustering algorithms and the real-time clustering algorithm, which includes the metaclustering algorithm.	40
2.13	Real-time clustering results with QrySim similarity measure that boosts the weights to the terms that occur close to the query term over the standard similarity measure (OrigSim) with equal weights to all terms.	40
2.14	Q_4 values with the real-time clustering algorithm with various combinations of features. The baselines include features with title and abstract and a single offline clustering algorithm. Although the combined feature set with all the features is the best one, the features with the offline clusters and title and abstract features are comparable to the ones that include body features.	41
2.15	Q_4 values with the real-time clustering algorithm and various granularity settings of offline clusters as features. The baseline feature set includes just title and abstract features. The numbers 1, 2, 3 refer to different settings of the offline clustering algorithm at different granularity settings, specifically varying from coarse to fine representation of clusters. It can be observed that the best accuracy is obtained by combining all the configurations, and individual cluster IDs themselves provide inferior performance.	41
3.1	Distribution of categories of medical concepts in EMERSE queries.	49
4.1	Classification and reranking.	77
4.2	Performance comparison between classification and ranking.	77
5.1	Relative feature importance in baseline.	90
5.2	Comparison of methods for modeling rating scores. Norm+Pred combines methods tagged using *. <i>b/z/m/c/q/r</i> indicates the significance over Baseline, ZeroOneNorm, MeanNorm, AutoNorm-C, AutoNorm-Q, and RatingPred, respectively, at the 0.001 level using the Wilcoxon nondirectional test.	93
5.3	Comparison of methods for modeling review counts. Norm+Pred combines methods tagged using *. The description of notations <i>b/z/m/c/q/r</i> are the same as Table 5.2 .	95
5.4	Comparison of methods for modeling distance. Methods with an indicator “+” apply logarithm transformation. <i>b/z/m/a</i> indicates the significance over Baseline, ZeroOneNorm+, Mean-Norm, and AutoNorm, respectively, at the 0.05 level using the Wilcoxon nondirectional test.	98
5.5	Comparison of methods for modeling user preference. Methods with an indicator “+” apply logarithm transformation. <i>b/n/m</i> indicates the significance over Baseline, NoNorm, and MeanNorm, respectively, at the 0.01 level using the Wilcoxon nondirectional test.	102
5.6	Sensitivity analysis. These data show that combining the proposed new features (i.e., All) can improve the Baseline over 7%.	102
5.7	Relative feature importance in the final model.	103
6.1	Example entity.	111
6.2	Example facet.	111
6.3	Features.	114

6.4	Entity popularity feature values of top entities.	116
6.5	CTR on category of the query entity vs. facet entity. Each row represents the category of the query entity and the column represents the category of the facet entity; each cell represents aggregate CTR at the intersection. The CTR values are normalized for each row such that the category with the highest CTR in each row is given 1.0. The missing entries indicate that the data for the intersection of those particular categories are not available.	123
6.6	Relevance improvements with various intercategory weights over the baseline. The smaller α , the more the intracategory relationships between facets are emphasized. DCG is computed for each group of facets with the same category.	124
6.7	DCG gain of various sets of features over the baseline.	125
7.1	The aspect relevance mapping function for local search.	135
7.2	Statistics of multi-aspect relevance data.	138
7.3	Statistics of overall relative preference datasets obtained through side-by-side comparison.	140
7.4	Evaluation of aggregation functions on label aggregation accuracy. Linear and Joint are significantly better than Rule (p-value < 001).	141
7.5	Evaluation of aggregation functions on ranking accuracy. Statistically significant differences (p-value < 001) compared to Rule are highlighted in bold.	142
9.1	Data summary for one source domain and three target domains.	193

List of Figures

About the Editors

Bo Long is currently a staff applied researcher at LinkedIn Inc., and was formerly a senior research scientist at Yahoo! Labs. His research interests lie in data mining and machine learning with applications to web search, recommendation, and social network analysis. He holds eight innovations patents and has published peer-reviewed papers in top conferences and journals including ICML, KDD, ICDM, AAAI, SDM, CIKM, and KAIS. He has served as reviewer, workshop co-organizer, conference organizer, committee member, and area chair for multiple conferences, including KDD, NIPS, SIGIR, ICML, SDM, CIKM, JSM, etc.

Yi Chang is a principal scientist and sciences director in Yahoo Labs, where he leads the search and anti-abuse science group. His research interests include web search, applied machine learning, natural language processing, and social computing. Yi has published more than 60 conference/journal papers and has served as workshop coorganizer, conference organizer, committee member, and area chair for multiple conferences, including WWW, SIGIR, ICML, KDD, CIKM, etc.

List of Contributors

Jaime Arguello University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Jiang Bian Microsoft Inc., Beijing, P.R. China

Yi Chang Yahoo! Labs, Sunnyvale, CA, USA

Fernando Diaz Microsoft Inc., New York, NY, USA

Anlei Dong Yahoo! Labs, Sunnyvale, CA, USA

David Hanauer University of Michigan, Ann Arbor, MI, USA

Yoshiyuki Inagaki Yahoo! Labs, Sunnyvale, CA, USA

Changsung Kang Yahoo! Labs, Sunnyvale, CA, USA

Yuan Liu Microsoft Inc., Beijing, P.R. China

Bo Long LinkedIn Inc., Mountain View, CA, USA

Yuanhua Lv Microsoft Inc., Mountain View, CA, USA

Qiaozhu Mei University of Michigan, Ann Arbor, MI, USA

Tao Mei Microsoft Inc., Beijing, P.R. China

Belle Tseng Apple Inc., Cupertino, CA, USA

Srinivas Vadrevu Microsoft Inc., Mountain View, CA, USA

Hongning Wang University of Illinois at Urbana-Champaign, Urbana, IL, USA

Xuanhui Wang Facebook Inc., Menlo Park, CA, USA

Kai Zheng University of Michigan, Ann Arbor, MI, USA

Foreword

As the information available on the Internet continues to grow, Web searchers increasingly run into a critical but not broadly discussed challenge: finding relevant, rich results for highly targeted and specialized queries.

General Web searching has been with us for a long time, has spawned international powerhouse companies like Yahoo! and Google, and is a staple of everyday life for hundreds of millions of people around the world. It is probable that literally billions of generic searches—for general information, celebrities, sports scores, common products, and other items of interest—are well satisfied by the commonly known major search engines whose names are so familiar as to have become part of the vernacular. But since the Internet has become the business- and daily-life critical worldwide resource it is today, increasingly diverse groups of people are relying on it to look for more and more diverse things—things not so easily found by generic Web search engines.

For example, searches focused on travel planning may often have very specific but implicit assumptions about results, such as the expectation of itineraries listed in order of departure or cost. They might also benefit from additional nonobvious information that could be critically helpful, such as changes in checked baggage policies, road construction near relevant airports, or State Department travel warnings. General search engines have no more clue about these things than they do about dosages of medications, celebrity divorces, slugging percentages, and other narrow, domain-specific information; they don't have access to site-specific signals. As a result, the role of so-called "vertical" search engines, which focus on specific segments of online content and deep site-specific information, has quietly increased to become essential to most people looking for key items online.

Somewhat hidden from view but no less important than general Web search technology, vertical search algorithms have been key to helping users find household products they care about, movies they want to see, potential dating partners, their perfect automobiles, well-matched insurance policies, and thousands of other things for which generalized text processing algorithms are not well tuned and cannot make the right judgments of relevance for results.

In this context, the term *vertical* is usually taken to connote in-depth treatment of fairly narrow domains, e.g., medical information, rather than broad ranges of information that meet a very wide range of needs (as you will see, for the purposes of this book, *vertical* can also refer to a limited range of search result types, such as entities or measurements or dates, or specific types of information access modalities, such as mobile search). What sets vertical search apart from more general, broad-based search is the fact that relatively specific domain knowledge can be leveraged to find the right pieces of information. Further, an understanding of a more limited set of information-seeking tasks (for example, looking for specific kinds of football statistics for your fantasy team) can also play an important role in satisfying narrower information needs. With fewer but very common tasks carried out by users, it may be easier to infer a user's intent for a particular vertical search, which could dramatically improve the quality and value of the results for the user.

In addition to the opportunity to provide highly relevant and richer results to information seekers

vertical search engines that focus on specific segments of online content have shown great potential to offer advertisers more contextually relevant, better-targeted audiences for their ads. Given the dependence of the Internet search industry on advertising, this makes vertical search an economical and central part of the Internet's future. There is no doubt that vertical search is starting to play a role in which the significance was probably never imagined in the early days of Internet searching.

As with other forms of search, the heart of successful vertical search is relevance ranking. Specialized understanding of the domain and sophisticated ranking algorithms is critical. Algorithms that work hard to infer a user's intention when doing a search are the ones that are successful. The ability to use the right signals and successfully compare various aspects of a query and potentially retrieved results will make or break a search engine. And that is the focus of this book: introducing and evaluating the critical ranking technology needed to make vertical searching successful.

Although there exist many books on general Web search technology, this new volume is a unique resource, dedicated to vertical search technologies and the relevance ranking technology that makes them successful. The book takes a comprehensive view of this area and aims to become an authoritative source of information for search scientists, engineers, and other interested readers with a technical bent. Despite many years of research on algorithms and methods of general Web search, vertical search deserves its own dedicated study and in-depth treatment because of the unique nature of its structures and applications. This volume provides that focused treatment, covering key issues such as cross-vertical searching, vertical selection and aggregation, news searches, object searches, image searches, and medical domain searches.

The authors represented in this book are active researchers who cover many different aspects of vertical search technology and who have made tangible contributions to the progress of what is clearly a dynamic research frontier. This ensures that the book is authoritative and reflects the current state of the art. Nevertheless—and importantly—the book gives a balanced treatment of a wide spectrum of topics, well beyond the individual authors' own methodologies and research specialties.

The book presents in-depth and systematic discussions of theories and practices for vertical search ranking. It covers the obvious major fields as well as recently emerging areas for vertical search, including news search ranking, local search ranking, object search ranking, image search ranking, medical domain search ranking, cross-vertical ranking, and vertical selection and aggregation. For each field, the book provides state-of-the-art algorithms with detailed discussions, including background, derivation, and comparisons. The book also presents extensive experimental results on various real application datasets to demonstrate the performance of various algorithms as well as guidelines for practical use of those algorithms. It introduces ranking algorithms for various vertical search ranking applications and teaches readers how to manipulate ranking algorithms to achieve better results in real applications. Finally, the book provides thorough theoretical analysis of various algorithms and problems to lay a solid foundation for future advances in the field.

Vertical search is still a fairly young and dynamic research field. This volume offers researchers and application developers a comprehensive overview of the general concepts, techniques, and applications of vertical search and helps them explore this exciting field and develop new methods and applications. It may also serve graduate students and other interested readers with a general introduction to the state of the art of this promising research area. It uses plain language with detailed examples, including case studies and real-world, hands-on examples to explain the key concepts, models, and algorithms used in vertical search ranking. I think that overall you will find it quite readable and highly informative.

Although not widely known, vertical search is an essential part of our everyday lives on the

Internet. It is increasingly critical to users' satisfaction and increasing reliance on online data sources and it provides extraordinary new opportunities for advertisers. Given recent growth in the application of vertical search and our increasing daily reliance on it, you hold in your hands the first guidebook to the next generation of information access on the Internet. I hope you enjoy it.

—Ron Brachman

Chief Scientist and Head, Yahoo! Labs

Yahoo!, Inc.

Introduction

Abstract

This book aims to present a systematic study of practices and theories for vertical search ranking. The studies in this book can be categorized into two major classes. One class is single-domain-related ranking that focuses on ranking for a specific vertical, such as news search ranking, medical domain search ranking, visual search ranking, mobile search ranking, and entity search ranking. Another class is multidomain-related ranking, which focuses on ranking that involves multiple verticals, such as multiaspect ranking, aggregating vertical search ranking, and cross-vertical ranking. This chapter discusses organization, audience, and further reading for this book.

Keywords

Vertical search ranking

news search ranking

medical domain search ranking

visual search ranking

mobile search ranking

multiaspect relevance ranking

entity ranking

aggregated vertical search

cross-vertical search ranking

1.1 Defining the Area

In the past decade, the impact of general Web search capabilities has been stunning. However, with exponential information growth on the Internet, it becomes more and more difficult for a general Web search engine to address the particular informational and research needs of niche users. As a response to the great need for deeper, more specific, more relevant search results, vertical search engines have emerged in various domains. By leveraging domain knowledge and focusing on specific user tasks, vertical search has great potential to serve users highly relevant search results from specific domains.

The core component of vertical search is *relevance ranking*, which has attracted more and more attention from both industry and academia during the past few years. This book aims to present a systematic study of practices and theories for vertical search ranking. The studies in this book can be categorized into two major classes. One class is single-domain-related ranking that focuses on ranking for a specific vertical, such as news search ranking and medical domain search ranking. However, in this book the term *vertical* has a more general meaning than topic. It refers to specific topics such as news and medical information, specific result types such as entities, and specific search

interfaces such as mobile search. The second class of vertical search study covered in this book class is multidomain-related ranking, which focuses on ranking involving multiple verticals, such as multispect ranking, aggregating vertical search ranking, and cross-vertical ranking.

1.2 The Content and Organization of This Book

This book aims to present an in-depth and systematic study of practices and theories related to vertical search ranking. The organization of this book is as follows.

Chapter 2 covers news vertical search ranking. News is one of the most important of Internet user online activities. For a commercial news search engine, it is critical to provide users with the most relevant and fresh ranking results. Furthermore, it is necessary to group the related news articles so that users can browse search results in terms of news stories rather than individual news articles. This chapter describes a few algorithms for news search engines, including ranking algorithms and clustering algorithms. For the ranking problem, the main challenge is achieving appropriate balance between topical relevance and freshness. For the clustering problem, the main challenge is how to group related news articles into clusters in a scalable mode. **Chapter 2** introduces a few news search ranking approaches, including a learning-to-rank approach and a joint learning approach from clickthroughs. The chapter then describes a scalable clustering approach to group news search results.

Chapter 3 studies another important vertical search, the medical domain search. With the exponential growth of electronic health records (EHRs), it is imperative to identify effective means to help medical clinicians as well as administrators and researchers retrieve information from EHRs. Recent research advances in natural language processing (NLP) have provided improved capabilities for automatically extracting concepts from narrative clinical documents. However, before these NLP-based tools become widely available and versatile enough to handle vaguely defined information retrieval needs by EHR users, a convenient and cost-effective solution continues to be in great demand. In this chapter, we introduce the concept of medical information retrieval, which provides medical professionals a handy tool to search among unstructured clinical narratives via an interface similar to that of general-purpose Web search engines, e.g., Google. In the latter part of the chapter, we also introduce several advanced features, such as intelligent, ontology-driven medical search query recommendation services and a collaborative search feature that encourages sharing of medical search knowledge among end users of EHR search tools.

Chapter 4 is intended to introduce some fundamental and practical technologies as well as some major emerging trends in visual search ranking. The chapter first describes the generic visual search system, in which three categories of visual search are presented: i.e., *text-based*, *query example-based*, and *concept-based* visual search ranking. Then we describe the three categories in detail, including a review of various popular algorithms. To further improve the performance of initial search results, visual search re-ranking of four paradigms will be presented: 1) *self-reranking*, which focuses on detecting relevant patterns from initial search results without any external knowledge; 2) *example-based reranking*, in which the query examples are provided by users so that the relevant patterns can be discovered from these examples; 3) *crowd-reranking*, which mines relevant patterns from crowd-sourcing information available on the Web; and 4) *interactive reranking*, which utilizes user interaction to guide the reranking process. In addition, we also discuss the relationship between machine learning and visual search, since most recent visual search ranking frameworks are developed based on machine learning technologies. Last, we conclude with several promising directions for future research.

Chapter 5 introduces *mobile search ranking*. The wide availability of Internet access on mobile devices, such as phones and personal media players, has allowed users to search and access Web information while on the go. The availability of continuous fine-grained location information on the devices has enabled mobile local search, which employs user location as a key factor to search for local entities (e.g., a restaurant, store, gas station, or attraction) to overtake a significant part of the query volume. This is also evident by the rising popularity of location-based search engines on mobile devices, such as Bing Local, Google Local, Yahoo! Local, and Yelp. The quality of any mobile local search engine is mainly determined by its ranking function, which formally specifies how we retrieve and rank local entities in response to a user's query. Acquiring effective ranking signals and heuristics to develop an effective ranking function is arguably the single most important research problem in mobile local search. This chapter first overviews the ranking signals in mobile local search (e.g., distance and customer rating score of a business), which have been recognized to be quite different from general Web search. We next present a recent data analysis that studies the behavior of mobile local search ranking signals using a large-scale query log, which reveals interesting heuristics that can be used to guide the exploitation of different signals to develop effective ranking features. Finally, we also discuss several interesting future research directions.

Chapter 6 is about *entity ranking*, which is a recent paradigm that refers to retrieving and ranking related objects and entities from different structured sources in various scenarios. Entities typically have associated categories and relationships with other entities. In this chapter, we introduce how to build a Web-scale entity ranking system based on machine-learning ranking models. Specifically, the entity ranking system usually takes advantage of structured knowledge bases, entity relationship graphs, and user data to derive useful features for facilitating semantic search with entities directly within the learning-to-rank framework. Similar to generic Web search ranking, entity pairwise preference can be leveraged to form the objective function of entity ranking. More than that, this chapter introduces ways to incorporate the categorization information and preference of related entities into the objective function for learning. This chapter further discusses how entity ranking is different from regular Web search in terms of presentation bias and the interaction of categories of query entities and result facets.

Chapter 7 presents learning to rank with multiaspect relevance for vertical searches. Many vertical searches, such as local search, focus on specific domains. The meaning of relevance in these vertical searches is domain-specific and usually consists of multiple well-defined aspects. For example, in local search, text matching and distance are two important aspects to assess relevance. Usually, the overall relevance between a query and a document is a tradeoff among multiple aspect relevancies. Given a single vertical, such a tradeoff can vary for different types of queries or in different contexts. In this chapter, we explore these vertical-specific aspects in the learning-to-rank setting. We propose a novel formulation in which the relevance between a query and a document is assessed with respect to each aspect, forming the multiaspect relevance. To compute a ranking function, we study two types of learning-based approaches to estimate the tradeoff among these aspect relevancies: a label aggregation method and a model aggregation method. Since there are only a few aspects, a minimum amount of training data is needed to learn the tradeoff. We conduct both offline and online bucket-test experiments on a local vertical search engine, and the experimental results show that our proposed multiaspect relevance formulation is very promising. The two types of aggregation methods perform more effectively than a set of baseline methods including a conventional learning-to-rank method.

Chapter 8 focuses on *aggregated vertical search*. Commercial information access providers increasingly incorporate content from a large number of specialized services created for particular

information-seeking tasks. For example, an aggregated Web search page may include results from image databases and news collections in addition to the traditional Web search results; a new provider may dynamically arrange related articles, photos, comments, or videos on a given article page. These auxiliary services, known as *verticals*, include search engines that focus on a particular domain (e.g., news, travel, sports), search engines that focus on a particular type of media (e.g., images, video, audio), and application programming interfaces (APIs) to highly targeted information (e.g., weather forecasts, map directions, or stock prices). The goal of *aggregated search* is to provide integrated access to all verticals within a single information context. Although aggregated search is related to classic work in distributed information retrieval, it has unique signals, techniques, and evaluation methods in the context of the Web and other production information access systems. In this chapter, we present the core problems associated with aggregated search, which include sources of predictive evidence, relevance modeling, and evaluation.

Chapter 9 presents recent advances in *cross-vertical ranking*. A traditional Web search engine conducts ranking mainly in a single domain, i.e., it focuses on one type of data source, and effective modeling relies on a sufficiently large number of labeled examples, which require an expensive and time-consuming labeling process. On the other side, it is very common for a vertical search engine to conduct ranking tasks in various verticals, which presents a more challenging ranking problem, that is *cross-domain ranking*. Although in this book our focus is on cross-vertical ranking, the proposed approaches can be applied to more general cases, such as cross-language ranking. Therefore, we use the more general term, cross-domain ranking, in this book. For cross-domain ranking, in some domains we may have a relatively large amount of training data, whereas in other domains we can only collect very little. Therefore, finding a way to leverage labeled information from related heterogeneous domains to improve ranking in a target domain has become a problem of great interest. In this chapter, we propose a novel probabilistic model, pairwise cross-domain factor (PCDF) model, to address this problem. The proposed model learns latent factors (features) for multidomain data in partially overlapped heterogeneous feature spaces. It is capable of learning homogeneous feature correlations, heterogeneous feature correlation, and pairwise preference correlation for cross-domain knowledge transfer. We also derive two PCDF variations to address two important special cases. Under the PCDF model, we derive a stochastic gradient-based algorithm, which facilitates distributed optimization and is flexible to adopt various loss functions and regularization functions to accommodate different data distributions. The extensive experiments on real-world data sets demonstrate the effectiveness of the proposed model and algorithm.

1.3 The Audience for This Book

The book covers major fields as well as recently emerging fields for vertical search. Therefore, the expected readership of this book includes all the researchers and systems development engineers working in these areas, including, but not limited to, Web search, information retrieval, data mining, and specific application areas related to vertical search, such as various specific vertical search engines. Since this book is self-contained in its presentation of the material, it also serves as an ideal reference book for people who are new to the topic of vertical search ranking. Consequently, in addition, the audience also includes anyone with interest or who works in a field requiring this reference book. Finally, this book can be used as a reference for a graduate course on advanced topics of information retrieval or data mining, since it provides a systematic introduction to this booming new subarea of information technology.

1.4 Further Reading

As a newly emerging area of information retrieval and data mining, vertical search ranking is still in its infant stage; currently there is no dedicated, premier venue for the publication of research in this area. Consequently, related work in this area, as the supplementary information to this book for further readings, may be found in the literature of the two parent areas.

In information retrieval, related work may be found in the premier conferences, such as the annual Association for Computing Machinery (ACM) Special Interest Group on Information Retrieval (SIGIR) conference, the International World Wide Web Conference (WWW), and the ACM, the International Conference on Information and Knowledge Management (ACM CIKM). For journals, the premier journals in the information retrieval area, including *Information Retrieval, Foundations and Trends in Information Retrieval* (FTIR), may contain related work in vertical search ranking.

In data mining, related work may be found in the premier conferences, such as the ACM International Conference on Knowledge Discovery and Data Mining (KDD), the Institute of Electrical and Electronics Engineers (IEEE), International Conference on Data Mining (ICDM), and the Society for Industrial and Applied Mathematics (SIAM) International Conference on Data Mining (SDM). In particular, related work may be found in the workshop dedicated to the area of relational learning, such as the Statistical Relational Learning workshop. The premier journals in the data mining area, including *IEEE Transactions on Knowledge and Data Engineering* (TKDE), *ACM Transactions on Data Mining* (TDM), and *Knowledge and Information Systems* (KAIS), may contain related work on relational data clustering.

News Search Ranking

Abstract

News search is one of the most important Internet user activities. For a commercial news search engine, it is critical to provide users with the most relevant and fresh ranking results. Furthermore, it is necessary to group the related news articles so that users can browse search results in terms of news stories rather than individual news articles. This chapter describes a few algorithms for news search engines, including ranking algorithms and clustering algorithms. For the ranking problem, the main challenge is achieving appropriate balance between topical relevance and freshness. For the clustering problem, the main challenge is grouping related news articles into clusters in a scalable mode. We begin by introducing a few news search ranking approaches including a learning-to-rank approach (Section 2.1) and a joint learning approach from clickthroughs (Section 2.2). We then describe a scalable clustering approach to group news search results (Section 2.3).

Keywords

News search

freshness

relevance

clustering

temporal features

2.1 The Learning-to-Rank Approach

The main challenge for ranking in news search is how to make appropriate balance between two factors: Relevance and freshness. Here relevance includes both topical relevance as well as news source authority.

A widely adopted approach in practice is to use a simple formula to combine relevance and freshness. For example, the final ranking score for a news article can be computed as

$$\text{score}_{rank} = \text{score}_{relevance} e^{-\beta t} \quad (2.1)$$

where $\text{score}_{relevance}$ is the value representing the relevance between query and news article, t is news article age and $e^{-\beta t}$ is a time decay term, for which the older a news article is, the more penalty the article will receive for its final ranking. The parameter β is used to control the relative importance of freshness in the final ranking result. In the literature of information retrieval, *document* is usually used to refer to a candidate item in ranking tasks. In this chapter, we use the terms *document* and *news article* equally because the application here is to rank news articles in a search.

The advantage of such a heuristic approach to a relevance and freshness combination is its efficiency in real practice, for which only the value of the parameter β needs to be tuned by using

some ranking examples. Furthermore, the appropriate β value often leads to good ranking results for many queries, which also makes this approach effective.

The drawback of this approach is that it is incapable of further improving ranking performance because such a heuristic rule is too naive to handle more complicated ranking cases. For example, (2.1), time decay is represented by the term $e^{-\beta t}$, which is totally dependent on the document age. In fact, an appropriate time decay factor should also rely on the nature of the query, since different queries have different time sensitivities: If a query is related to breaking news, such as an earthquake that has just happened and has extensive media reports on casualty and rescue, then freshness should be very important because even a document published only one hour ago could be outdated. On the other hand, if a query is for an event that happened weeks ago, then relevance is more important for ranking because the user would like to find the most relevant and comprehensive reports in the search results.

2.1.1 Related Works

Many prior works have exploited the temporal dimension in searches. For example, Baeza-Yates *et al.* [22] studied the relation among Web dynamics, structure, and page quality and demonstrated that PageRank is biased against new pages. In T-Rank Light and T-Rank algorithms [25], both activity (i.e., update rates) and freshness (i.e., timestamps of most recent updates) of pages and links are taken into account in link analysis. Cho *et al.* [66] proposed a page quality ranking function in order to alleviate the problem of popularity-based ranking, and they used the derivatives of PageRank to forecast future PageRank values for new pages. Nunes [269] proposed to improve Web information retrieval in the temporal dimension by combining the temporal features extracted from both individual documents and the whole Web. Pandey *et al.* [276] studied the tradeoff between new page exploration and high-quality page exploitation, which is based on a ranking method to randomly promote some new pages so that they can accumulate links quickly.

Temporal dimension is also considered in other information retrieval applications. Del Corso *et al.* [94] proposed the ranking framework to model news article generation, topic clustering, and story evolution over time, and this ranking algorithm takes publication time and linkage time into consideration as well as news source authority. Li *et al.* [221] proposed a TS-Rank algorithm, which considers page freshness in the stationary probability distribution of Markov chains, since the dynamics of Web pages are also important for ranking. This method proves effective in the application of publication search. Pasca [277] used temporal expressions to improve question answering results for time-related questions. Answers are obtained by aggregating matching pieces of information and the temporal expressions they contain. Furthermore, Arikan *et al.* [20] incorporated temporal expressions into a language model and demonstrated experimental improvement in retrieval effectiveness.

Recency query classification plays an important role in recency ranking. Diaz [98] determined the newsworthiness of a query by predicting the probability of a user clicking on the news display of a query. König *et al.* [204] estimated the clickthrough rate for dedicated news search results with a supervised model, which is to satisfy the requirement of adapting quickly to emerging news event.

2.1.2 Combine Relevance and Freshness

Learning-to-rank algorithms have shown significant and consistent success in various applications.

[226,184,406,54]. Such machine-learned ranking algorithms learn a ranking mechanism by optimizing particular loss functions based on editorial annotations. An important assumption in those learning methods is that document relevance for a given query is generally stationary over time, so that, long as the coverage of the labeled data is broad enough, the learned ranking functions would generalize well to future unseen data. Such an assumption is often true in Web searches, but it is less likely to hold in news searches because of the dynamic nature of news events and the lack of time annotations.

A typical procedure is as follows:

- Collect query-URL pairs.
- Ask editors to label the query-URL pairs with relevance grades.
- Apply a learning-to-rank algorithm to the train ranking model.

Traditionally, in learning-to-rank, editors label query-URL pairs with relevance grades, which usually have four or five values, including *perfect*, *excellent*, *good*, *fair*, or *bad*. Editorial labeling information is used for ranking model training and ranking model evaluation. For training, the relevance grades are directly mapped to numeric values as learning targets.

For evaluation, we desire an evaluation metric that supports graded judgments and penalizes errors near the beginning of the ranked list. In this work, we use DCG [175],

$$\text{DCG}_n = \sum_{i=1}^n \frac{G_i}{\log_2(i+1)}, \quad (2.2)$$

where i is the position in the document list, and G_i is the function of relevance grade. Because the range of DCG values is not consistent across queries, we adopt the NDCG as our primary ranking metric,

$$\text{NDCG}_n = Z_n \sum_{i=1}^n \frac{G_i}{\log_2(i+1)}, \quad (2.3)$$

where Z_n is a normalization factor, which is used to make the NDCG of the ideal list be 1. We can use NDCG_1 and NDCG_5 to evaluate the ranking results.

We extend the learning-to-rank algorithm in news searches, for which we mainly make two modifications due to the dynamic nature of the news search: (1) training sample collection and (2) editorial labeling guideline.

2.1.2.1 Training Sample Collection

The training sample collection has to be near real time for news searches by the following steps:

1. Sample the *latest* queries from the news search query log.
2. *Immediately* get the candidate URLs for the sampled queries.
3. *Immediately* ask editors to do judgments on the query-URL pairs with relevance and freshness grades.

We can see that all the steps need to be accomplished in a short period. Therefore, the training sample collection has to be well planned in advance; otherwise, any delay during this procedure would

affect the reliability of the collected data. If queries are sampled from an outdated query log or if a portion of the selected candidate URLs are outdated, they cannot represent the real data distribution. If editors do not label query-URL pairs on time, it will be difficult for them to provide accurate judgments because editors' judgments rely on their good understanding of the related news events, which becomes more difficult as time elapses.

2.1.2.2 Editorial Labeling

In a news search, editors should provide query-URL grades on both traditional relevance and freshness. Although document age is usually available in news searches, it is impossible to determine a document's freshness based solely on document age. A news article published one day ago could either be very fresh or very old, depending on the nature of the related news event. So, we ask editors to provide subjective judgments on document freshness, which can have a few different grades, such as the following:

- Very fresh: latest documents (promote grade: 1)
- Fresh (promote grade: 0)
- A little bit outdated (demote grade: -1)
- Totally outdated and useless (demote grade: -2)

For ranking model training, we combine relevance grade and recency grade for a new grade as a learning target. An example of such a grade combination is shown in the list. If the document is very fresh, we promote one grade from its original relevance grade. For example, if a document has a *good* grade on relevance and a *fresh* grade on freshness, its final grade is *excellent*, which is one grade higher than its relevance grade, *good*. If the document is fresh, we neither promote nor demote. If the document is a little bit outdated, we demote one grade from its original relevance grade. If the document is totally outdated and useless, we demote two grades.

For ranking model evaluation, we can compute DCG values based on either the combined grades or the relevance grade and freshness grade separately.

To evaluate freshness in isolation, we also include a freshness metric based on DCG, called DCF:

$$\text{DCF}_n = \sum_{i=1}^n \frac{F_i}{\log_2(i+1)}, \quad (2.4)$$

where i is the position in the document list, and F_i is the freshness label (1 or 0). A query may have multiple very fresh documents—for example, when multiple news sources simultaneously publish updates to some ongoing news story. Note that DCF is a recency measurement that is independent of overall relevance. Therefore, when we evaluate a ranking, we should first consider demoted NDCG, which represents the overall relevance, then inspect the value of the DCF. We also define normalized discounted cumulative freshness (NDCF) in the same way as for NDCG in Eq. 2.3.

2.2 Joint Learning Approach from Clickthroughs

Given the highly dynamic nature of news events and the sheer scale of news reported around the globe, it is often impractical for human editors to constantly keep track of the news and provide timely relevance and freshness judgments. To validate this conjecture, we asked editors to annotate

sample content of Relevance Ranking for Vertical Search Engines

- [read Sociolinguistics: A Very Short Introduction pdf, azw \(kindle\)](#)
- [American Gun: A History of the U.S. in Ten Firearms pdf, azw \(kindle\)](#)
- [download Stallo \(UK Edition\) pdf, azw \(kindle\)](#)
- [Sugar: A Global History \(Edible Series\) online](#)
- [download online Collected Plays, Volume 1](#)
- [download online The Nature of Consciousness: Philosophical Debates pdf, azw \(kindle\), epub, doc, mobi](#)

- <http://korplast.gr/lib/Enterprise-Android--Programming-Android-Database-Applications-for-the-Enterprise.pdf>
- <http://www.experienceolvera.co.uk/library/American-Gun--A-History-of-the-U-S--in-Ten-Firearms.pdf>
- <http://academialanguagebar.com/?ebooks/Stallo--UK-Edition-.pdf>
- <http://berttrotman.com/library/Sugar--A-Global-History--Edible-Series-.pdf>
- <http://growingsomeroots.com/ebooks/Collected-Plays--Volume-1.pdf>
- <http://schrofl.de/books/The-Nature-of-Consciousness--Philosophical-Debates.pdf>