Adaptation, Learning, and Optimization 13

Series Editors-in-Chief

Meng-Hiot Lim
Division of Circuits and Systems
School of Electrical & Electronic Engineering
Nanyang Technological University
Singapore 639798
E-mail: emhlim@ntu.edu.sg

Yew-Soon Ong
School of Computer Engineering
Nanyang Technological University
Block N4, 2b-39
Nanyang Avenue
Singapore 639798
E-mail: asysong@ntu.edu.sg

For further volumes:
http://www.springer.com/series/8335

Shengli Wu

Data Fusion in Information Retrieval

Springer

Author
Shengli Wu
University of Ulster
United Kingdom

ISSN 1867-4534 e-ISSN 1867-4542
ISBN 978-3-642-28865-4 e-ISBN 978-3-642-28866-1
DOI 10.1007/978-3-642-28866-1
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012935112

© Springer-Verlag Berlin Heidelberg 2012
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

These days the technique of data fusion has been used extensively in many different application areas and information retrieval is one of them. As early as 1980's, researchers began to investigate combining retrieval results generated from many different ways. Since then the problem of data fusion has been investigated or applied by many researchers in the information retrieval community. Looking at the technical implementation of information retrieval systems, one may find that data fusion has been used by many of them, if not all of them, in one way or another. This demonstrates that data fusion, as a useful technique for performance improvement, has been accepted by the information retrieval community. But sometimes, the best possible fusion technique has not been used.

In 2001 I took part in an EU Project that was relevant to data fusion. I published my first paper on data fusion in 2002. Since then, I have been spending most of my research time on data fusion and some related topics in information retrieval. It has been 10 years and I have done a lot of research work on it. I think it is a good idea to put all my research output together in a systematic style so as to make it easy for others to find relevant material. The book should be useful for researchers and students in the areas of information retrieval, web search, digital libraries, and online information service, to obtain the up-to-date data fusion technique. The book may also be useful for people in other research areas such as machine learning, artificial neural networks, classification of text documents or/and multimedia files, and so on, since the data fusion technique has been widely used in those areas as well.

The book has a good coverage of different aspects of the data fusion technique: from retrieval evaluation (Chapter 2) to score normalization (Chapter 3); from score-based methods (Chapters 4, 5, and 6) to ranking-based methods (Chapter 7); from theoretical framework (Chapter 6) to data fusion applications (Chapter 9). Experiments have been conducted and reported for almost all of the aspects discussed. A related problem, fusing results from overlapping databases, is also addressed (Chapter 8).

Information retrieval is an area that requires effort on both theoretical and empirical sides. There is no exception for the data fusion problem. The geometric framework discussed in Chapter 6 provides us a good platform to understand the

properties of (score-based) data fusion algorithms; while results of various experiments help us to find useful information such as the relation between rank and relevance in a resultant list, or the performance of a given data fusion algorithm in various kinds of situations.

Apart from some examples of application in Chapter 9, I wrote programs for almost all those empirical studies included in this book. Sometimes it is a bitter experience when I found some results obtained were inconsistent or suspicious, but could not find where the things might go wrong. But after so many data fusion experiments, I am now happy to see that in general, the results are consistent and convincing across different experiments. For example, CombSum, CombMNZ, the Borda count, the linear combination method have been tested in different data sets for many times, and the experimental results are very consistent.

Finally, I would give my thanks to my colleagues and friends, especially Fabio Crestani, Sally McClean, Yaxin Bi, Xiaoqin Zeng, and others, who worked with me on the data fusion problem or helped me somehow; I would give my thanks to my wife, Jieyu Li, who always backs me; and I would give my thanks to my son, Zao Wu, who proof-read most of my published papers and the manuscript of this book as well.

Shengli Wu
University of Ulster
Jordanstown, Northern Ireleand, UK
January, 2012

Contents

1 Introduction 1
1.1 The Rationale Behind Data Fusion 1
1.2 Several Typical Data Fusion Methods 2
1.3 Several Issues in Data Fusion 4

2 Evaluation of Retrieval Results 7
2.1 Binary Relevance Judgment 7
2.2 Incomplete Relevance Judgment 10
2.3 Graded Relevance Judgment 11
2.4 Score-Based Metrics 14
2.5 Uncertainty of Data Fusion Methods on Effectiveness Improvement 18

3 Score Normalization 19
3.1 Linear Score Normalization Methods 20
3.1.1 The Zero-One Linear Method 20
3.1.2 The Fitting Method 21
3.1.3 Normalizing Scores over a Group of Queries 22
3.1.4 Z-Scores 22
3.1.5 Sum-to-One 23
3.1.6 Comparison of Four Methods 24
3.2 Nonlinear Score Normalization Methods 27
3.2.1 The Normal-Exponential Mixture Model 27
3.2.2 The CDF-Based Method 27
3.2.3 Bayesian Inference-Based Method 28
3.3 Transforming Ranking Information into Scores 29
3.3.1 Observing from the Data 29
3.3.2 The Borda Count Model 30
3.3.3 The Reciprocal Function of Rank 31
3.3.4 The Logistic Model 31

3.3.5 The Cubic Model 32
3.3.6 The Informetric Distribution 33
3.3.7 Empirical Investigation 35
3.4 Mixed Normalization Methods 40
3.5 Related Issues 41
3.5.1 Weights Assignment and Normalization 41
3.5.2 Other Relevance Judgments 41

4 Observations and Analyses 43
4.1 Prior Observations and Points of View 43
4.2 Performance Prediction of CombSum and CombMNZ 46
4.2.1 Data Fusion Performance 48
4.2.2 Improvement over Average Performance 51
4.2.3 Performance Improvement over Best Performance 52
4.2.4 Performance Prediction 54
4.2.5 The Predictive Ability of Variables 55
4.2.6 Some Further Observations 57
4.2.7 Summary 61
4.3 Comparison of CombSum and CombMNZ 62
4.3.1 Applying Statistical Principles to Data Fusion 63
4.3.2 Further Discussion about CombSum and CombMNZ 64
4.3.3 Lee's Experiment Revisited 66
4.4 Performance Prediction of the Linear Combination Method 70

5 The Linear Combination Method 73
5.1 Performance Related Weighting 73
5.1.1 Empirical Investigation of Appropriate Weights 75
5.1.2 Other Features 81
5.2 Consideration of Uneven Similarity among Results 87
5.2.1 Correlation Methods 1 and 2 87
5.2.2 Correlation Methods 3 and 4 88
5.2.3 Correlation Methods 5 and 6 89
5.2.4 Empirical Study 91
5.2.5 Some More Observations 94
5.3 Combined Weights 97
5.3.1 Theory of Stratified Sampling and Its Application 97
5.3.2 Experiments 99
5.4 Deciding Weights by Multiple Linear Regression 102
5.4.1 Empirical Investigation Setting 103
5.4.2 Experiment 1 104
5.4.3 Experiment 2 109
5.4.4 Experiment 3 111
5.5 Optimization Methods 114
5.6 Summary 116

6 A Geometric Framework for Data Fusion 117
6.1 Relevance Score-Based Data Fusion Methods and Ranking-Based Metrics 117
6.2 The Geometric Framework 119
6.3 The Centroid-Based Data Fusion Method 120
6.4 The Linear Combination Method 126
6.5 Relation between the Euclidean Distance and Ranking-Based Measures 129
6.6 Conclusive Remarks 132

7 Ranking-Based Fusion 135
7.1 Borda Count and Condorcet Voting 135
7.2 Weights Training for Weighted Condorcet Voting 137
7.3 Experimental Settings and Methodologies 139
7.4 Experimental Results of Fusion Performance 140
7.5 Positive Evidence in Support of the Hypothesis 144

8 Fusing Results from Overlapping Databases 149
8.1 Introduction to Federated Search 149
8.2 Resource Selection 152
8.2.1 A Basic Resource Selection Model 152
8.2.2 Solution to the Basic Resource Selection Model 156
8.2.3 Some Variants of the Basic Model and Related Solutions . . . 159
8.2.4 Experiments and Experimental Results 163
8.2.5 Further Discussion 165
8.3 Results Merging 166
8.3.1 Several Result Merging Methods 167
8.3.2 Evaluation of Result Merging Methods 169
8.3.3 Overlap between Different Results 177
8.3.4 Conclusive Remarks 178

9 Application of the Data Fusion Technique 181
9.1 Ranking Information Retrieval Systems with Partial Relevance Judgment 181
9.1.1 How Is the Problem Raised? 182
9.1.2 Other Options Than Averaging All the Values for Ranking Retrieval Systems 185
9.1.3 Evaluation of the Four Ranking Methods 186
9.1.4 Remarks 189
9.2 Ranking Information Retrieval Systems without Relevance Judgment 189
9.2.1 Methodology 191
9.2.2 Experimental Settings and Results 192

9.3 Applying the Data Fusion Technique to Blog Opinion Retrieval 193
9.3.1 Data Set and Experimental Settings 194
9.3.2 Experimental Results 195
9.3.3 Discussion and Further Analysis 202
9.4 TREC Systems via Combining Features and/or Components 203
9.4.1 A Hierarchical Relevance Retrieval Model for Entity Ranking 203
9.4.2 Combination of Evidence for Effective Web Search 205
9.4.3 Combining Candidate and Document Models for Expert Search 206
9.4.4 Combining Lexicon-Based Methods to Detect Opinionated Blogs 207
9.4.5 Combining Resources to Find Answers to Biomedical Questions 210
9.4.6 Access to Legal Documents: Exact Match, Best Match, and Combinations 210
9.4.7 Summary 212

References 213

A Systems Used in Section 5.1 221

B Systems Used in Section 5.2.4 225

C Systems Used in Sections 7.3-7.5 227

Notation used throughout this book (1)

Symbol	Meaning
CombMNZ	a data fusion method
CombSum	a data fusion method
D	document collection
$\|D\|$	number of documents in D
d_i	a document in collection D
$dist(A,B)$	Euclidean distance between A and B
$g(d_i)$	global score of document d_i, used for ranking the fused result
IR	a group of information retrieval systems
ir_i	an information retrieval system
LC	linear combination, a data fusion method
LCP	the LC method with performance-level weighting
LCP2	the LC method with performance square weighting
LCR	the LC method with weights trained by multiple regression
L_i	a result (a ranked list of documents) from ir_i for given D and q
$\|L_i\|$	number of documents in L_i
o_rate	overlap rate between two or more retrieval results
p_i	performance of a retrieval system ir_i over one or a group of queries
$Pr(t)$	probability of a document at rank t being relevant
Q	a group of queries
$\|Q\|$	number of queries in Q
q or q_i	a query in Q
r_i	raw score of document d_i
s_i	normalized score of document d_i
t_i	ranking position of document d_i
w_i	weight assigned to ir_i

Notation used throughout this book (2)

Metric	Meaning
AP	average precision over all relevant documents, or average precision for short
APD	average precision over all documents
bpref	binary preference
DCG	discounted cumulative gain
infAP	inferred average precision
NAPD	normalized average precision over all documents
NDCG	normalized discounted cumulative gain
NDCG@k	NDCG that sets 1 to be the weight of the first k documents
P@x	precision at x document-level, a metric for retrieval evaluation
PIA	percentage of improvement of the fused result over average of all component results
PIA(AP)	PIA measured by AP
PIA(RP)	PIA measured by RP
PIA(RR)	PIA measured by RR
PIB	percentage of improvement of the fused result over the best component result
PIB(AP)	PIB measured by AP
PIB(RP)	PIB measured by RP
PIB(RR)	PIB measured by RR
PFB	percentage of runs that the fused result is better than the best component result
PFB(AP)	PFB measured by AP
PFB(RP)	PFB measured by RP
PFB(RR)	PFB measured by RR
RP	recall-level precision
RR	reciprocal rank