Federated Search

Full text available at: http://dx.doi.org/10.1561/1500000010

Federated Search

Milad Shokouhi

Microsoft Research
Cambridge, CB30FB
UK
milads@microsoft.com

Luo Si

Purdue University
West Lafayette, IN 47907-2066
USA
lsi@cs.purdue.edu

now

the essence of knowledge

Boston – Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is M. Shokouhi and L. Si, Federated Search, Foundation and Trends® in Information Retrieval, vol 5, no 1, pp 1–102, 2011

ISBN: 978-1-60198-422-7

Foundations and Trends® in Information Retrieval
Volume 5 Issue 1, 2011

Editorial Board

Editorial Scope

Foundations and Trends® in Information Retrieval will publish survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians
Foundations and Trends® in Information Retrieval, 2011, Volume 5, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in
Information Retrieval
Vol. 5, No. 1 (2011) 1–102

DOI: 10.1561/1500000010

now
the essence of knowledge

Federated Search

Milad Shokouhi[1] and Luo Si[2]

[1] *Microsoft Research, 7 JJ Thomson Avenue, Cambridge, CB30FB, UK, milads@microsoft.com*
[2] *Purdue University, 250N University Street, West Lafayette, IN 47907-2066, USA, lsi@cs.purdue.edu*

Abstract

Federated search (federated information retrieval or distributed information retrieval) is a technique for searching multiple text collections simultaneously. Queries are submitted to a subset of collections that are most likely to return relevant answers. The results returned by selected collections are integrated and merged into a single list. Federated search is preferred over centralized search alternatives in many environments. For example, commercial search engines such as Google cannot easily index uncrawlable hidden web collections while federated search systems can search the contents of hidden web collections without crawling. In enterprise environments, where each organization maintains an independent search engine, federated search techniques can provide parallel search over multiple collections.

There are three major challenges in federated search. For each query, a subset of collections that are most likely to return relevant documents are selected. This creates the *collection selection* problem. To be able to select suitable collections, federated search systems need to acquire some knowledge about the contents of each collection, creating

the *collection representation* problem. The results returned from the selected collections are merged before the final presentation to the user. This final step is the *result merging* problem.

The goal of this work, is to provide a comprehensive summary of the previous research on the federated search challenges described above.

Contents

1

Introduction

Internet search is one of the most popular activities on the web. More than 80% of internet searchers use search engines for finding their information needs [251]. In September 1999, Google claimed that it received 3.5 million queries per day.[1] This number increased to 100 million in 2000,[2] and has grown to hundreds of millions since.[3] The rapid increase in the number of users, web documents and web queries shows the necessity of an advanced search system that can satisfy users' information needs both effectively and efficiently.

Since Aliweb [150] was released as the first internet search engine in 1994, searching methods have been an active area of research, and search technology has attracted significant attention from industrial and commercial organizations. Of course, the domain for search is not limited to the internet activities. A person may utilize search systems to find an email in a mail box, to look for an image on a local machine, or to find a text document on a local area network.

[1] http://www.google.com/press/pressrel/pressrelease4.html, accessed on 17 Aug 2010.
[2] http://www.google.com/corporate/history.html, accessed on 17 Aug 2010.
[3] http://www.comscore.com/Press_Events/Press_Releases/2010/8/comScore_Releases_July_2010_U.S._Search_Engine_Rankings, accessed on 17 Aug 2010.

Commercial search engines use programs called crawlers (or spiders) to download web documents. Any document overlooked by crawlers may affect the users perception of what information is available on the web. Unfortunately, search engines cannot easily crawl documents located in what is generally known as the *hidden web* (or *deep web*) [206]. There are several factors that make documents uncrawlable. For example, page servers may be too slow, or many pages might be prohibited by the robot exclusion protocol and authorization settings. Another reason might be that some documents are not linked to from any other page on the web. Furthermore, there are many *dynamic pages* — pages whose content is generated on the fly — that are crawlable [206] but are not bounded in number, and are therefore often ignored by crawlers.

As the size of the hidden web has been estimated to be many times larger than the number of visible documents on the web [28], the volume of information being ignored by search engines is significant. Hidden web documents have diverse topics and are written in different languages. For example, PubMed[4] — a service of the US national library of medicine — contains more than 20 million records of life sciences and biomedical articles published since the 1950s. The US census Bureau[5] includes statistics about population, business owners and so on in the USA. There are many patent offices whose portals provide access to patent information, and there are many other websites such for yellow pages and white pages that provide access to hidden web information.

Instead of expending effort to crawl such collections — some of which may not be crawlable at all — *federated search* techniques directly pass the query to the search interface of suitable collections and merge their results. In federated search, queries are submitted directly to a set of searchable collections — such as those mentioned for the hidden web — that are usually distributed across several locations. The final results are often comprised of answers returned from multiple collections.

[4] `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed`, accessed on 17 Aug 2010.
[5] `http://www.census.gov`, accessed on 17 Aug 2010.

From the users' perspective, queries should be executed on servers that contain the most relevant information. For example, a government portal may consist of several searchable collections for different organizations and agencies. For a query such as 'Administrative Office of the US Courts', it might not be useful to search all collections. A better alternative may be to search only collections from the `www.uscourts.gov` domain that are likely to contain the relevant answers.

However, federated search techniques are not limited to the web and can be useful for many enterprise search systems. Any organization with multiple searchable collections can apply federated search techniques. For instance, Westlaw[6] provides federated search for legal professionals covering more than 30,000 databases [59, 60, 61]. The users can search for case law, court documents, related newspapers and magazines, public records, and in return, receive merged results from heterogeneous sources. FedStats[7] is an online portal of statistical information published by many federal agencies. The crawls for the original centralized search in FedStats could be updated only every three months. Therefore, a federated search solution was requested and this was the main focus of the FedLemur project [13].[8] FedStats enables citizens, businesses, and government employees to find useful information without separately visiting web sites of individual agencies.

Federated search can be also used for searching multiple catalogs and other information sources. For example, in the Cheshire project,[9] many digital libraries including the UC Berkeley Physical Sciences Libraries, Penn State University, Duke University, Carnegie Mellon University, UNC Chapel Hill, the Hong Kong University of Science and Technology and a few other libraries have become searchable through a single interface at the University of Berkeley. Similarly, The European Library[10] provides a federated search solution to access the resources of 47 national libraries.

---

[6] `http://www.thomsonreuters.com/products_services/legal/legal_products/393832/Westlaw`, accessed on 17 Aug 2010.
[7] `http://search.fedstats.gov`, accessed on 17 Aug 2010.
[8] FedStats search is currently powered by `google.com`.
[9] `http://cheshire.berkeley.edu/`, accessed on 17 Aug 2010.
[10] `http://search.theeuropeanlibrary.org/portal/en/index.html`, accessed on 17 Aug 2010.

1.1 Federated Search

In federated search systems,[11] the task is to search a group of independent collections, and to effectively merge the results they return for queries.

Figure 1.1 shows the architecture of a typical federated search system. A central section (the *broker*) receives queries from the users and sends them to collections that are deemed most likely to contain relevant answers. The highlighted collections in Figure 1.1 are those selected for the query. To route queries to suitable collections, the broker needs to store some important information (summary or representation) about available collections. In a *cooperative* environment, collections inform brokers about their contents by providing

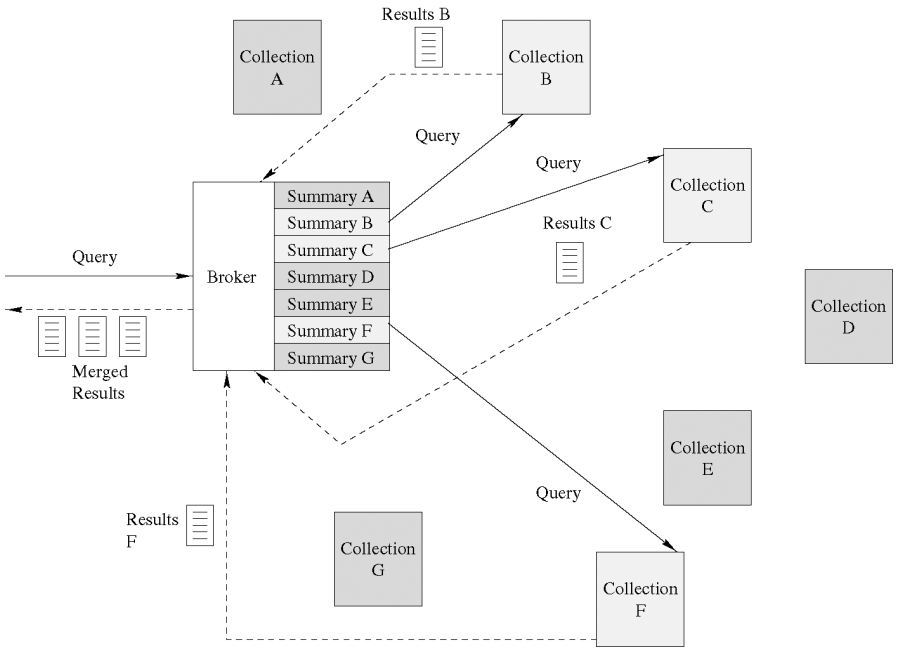

Fig. 1.1 The architecture of a typical federated search system. The broker stores the representation set (the summary) of each collection, and selects a subset of collections for the query. The selected collections then run the query and return their results to the broker, which merges all results and ranks them in a single list.

[11] Also referred to as distributed information retrieval (DIR).

information such as their term statistics. This information is often exchanged through a set of shared protocols such as STARTS [111] and may contain term statistics and other metadata such as collection size. In *uncooperative* environments, collections do not provide any information about their contents to brokers. A technique that can be used to obtain information about collections in such environments is to send sampling (*probe*) queries to each collection. Information gathered from the limited number of answer documents that a collection provides in response to such queries is used to construct a *representation set*; this representation set guides the evaluation of user queries and ranking collections. The selected collections receive the query from the broker and evaluate it on their own indexes. In the final step, the broker ranks the results returned by the selected collections and presents them to the user.

Federated search systems therefore need to address three major issues: how to represent the collections, how to select suitable collections for searching; and how to merge the results returned from collections.[12] Brokers typically compare each query to representation sets — also called summaries [138] — of each collection, and estimate the goodness of the collection accordingly. Each representation set may contain statistics about the lexicon of the corresponding collection. If the lexicon of the collections is provided to the central broker — that is, if the collections are cooperative — then complete and accurate information can be used for collection selection. However, in an uncooperative environment such as the hidden web, the collections need to be sampled to establish a summary of their topic coverage. This technique is known as query-based sampling [42] or query probing [116].

Once the collection summaries are generated, the broker has sufficient knowledge for collection selection. It is usually not feasible to search all collections for a query due to time constraints and bandwidth restrictions. Therefore, the broker selects a few collections that are most likely to return relevant documents based on their summaries. The selected collections receive the query and return their results to the broker.

[12] We briefly describe other common challenges such as *building wrappers* in Section 2.

Result merging is the last step of a federated search session. The results returned by multiple collections are gathered and ranked by the broker before presentation to the user. Since documents are returned from collections with different lexicon statistics and ranking features, their scores or ranks are not comparable. The main goal of result merging techniques is computing comparable scores for documents returned from different collections, and ranking them accordingly.

1.2 Federated Search on the Web

The most common forms of federated search on the web include *vertical* search, *peer-to-peer* (P2P) networks, and *metasearch* engines. Vertical search — also known as aggregated search — blends the top-ranked answers from search verticals (e.g., images, videos, maps) into the web search results. P2P search connects distributed peers (usually for file sharing), where each peer can be both *server* and *client* . Metasearch engines combine the results of different search engines in single results lists. Depending on the query, metasearch engines can select different engines for blending.

1.2.1 Vertical (aggregated) Search

Until recently, web search engines used to only show text answers in their results. Users interested in other types of answers (e.g., images, videos, and maps), had to directly submit their queries to the specialized *verticals*.

In 2000, the Korean search engine Naver[13] introduced *comprehensive search* and blended multimedia answers in their default search results. In May 2007, Google launched aggregated search (universal search) "to break down the walls that traditionally separated [their] various search properties and integrate the vast amounts of information available into one simple set of search results".[14] In aggregated search, the top-ranked answers from other information sources (e.g., image vertical) are merged with the default text results. Universal

[13] `http://www.naver.com`, accessed on 17 Aug 2010.

[14] `http://googleblog.blogspot.com/2007/05/universal-search-best-answer-is-still.html`, accessed on 17 Aug 2010.

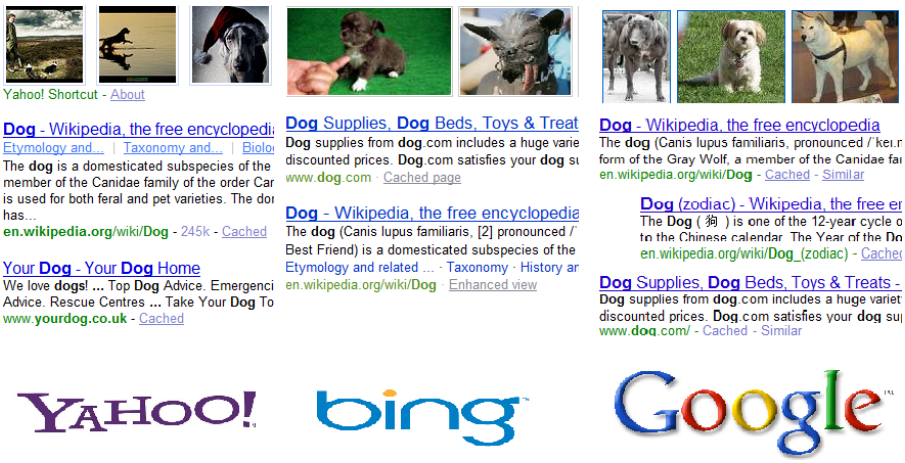

Fig. 1.2 The outputs of three major search engines for the query "dog". The top-ranked answers from the image vertical are blended in the final results.

search substantially increased the traffic of Google's non-text search verticals. For instance, the traffic of Google Maps increased by more than 20%.[15] Since then, all other major search engines such as Yahoo![16] and Bing have adopted aggregated search techniques. Figure 1.2 shows the results returned by three major search engines for the query "dog". It can be seen that all search engines merge some image answers along with their text results.

An aggregated search interaction consists of two major steps: *vertical selection* and *merging*. In the first step, the verticals relevant to the query are selected. A few examples of common verticals that are utilized by current search engines are: images, videos, news, maps, blogs, groups and books. The answers returned from the selected verticals are integrated with the default web results in the *merging* step.

Aggregated search was discussed in a workshop at SIGIR 2008 [191] as a promising area of research. Less than a year after, Diaz [77] proposed a click-based classifier for integration of news answers into web search results — as the first large-scale published study on aggregated

[15] `http://searchengineland.com/070608-091826.php`, accessed on 17 Aug 2010.
[16] Yahoo! has recently launched a new website (`http://au.alpha.yahoo.com/`) that applies aggregated search on a greater number of data sources.

search that won the best paper award at WSDM 2009.[17] Arguello et al. [9] proposed a classification-based method for vertical selection. The authors trained a classifier with features derived from the query string, previous query logs, and vertical content. They tested their techniques on a framework of 18 verticals, for which they won the best paper award at SIGIR 2009.[18] Diaz and Arguello [78] showed that integrating users feedback such as clicks can significantly improve the performance of vertical selection methods.

Aggregated search is a new area of research, and has opened several directions for future work; what search verticals shall be selected for a query? How can the results of different verticals be merged into a single list? Do users prefer aggregated search results? How aggregated search changes users' search behaviors?

1.2.2 Peer-to-peer Networks

Lu [168] showed that the search task in a peer-to-peer network is closely related with the research topic of federated search. A peer-to-peer network (P2P) consists of three main types of objects: information providers, information consumers, and a search mechanism that retrieves relevant information from providers for consumers.

The P2P network architectures can be divided into four categories: *broker-based* P2P networks (e.g., the original Napster music file-sharing system[19]) have a single centralized service that also contains document lists shared from peer nodes. The centralized service responds to queries from consumers by returning the pointers of relevant documents. In *Decentralized* P2P architectures such as Gnutella v0.4[20] each peer node can serve as both provider and consumer. *Hierarchical* P2P architectures such as, Gnutella v0.6[21], Gnutella2[22], BearShare[23] and

[17] `http://www.wsdm2009.org`, accessed on 17 Aug 2010.
[18] `http://sigir2009.org`, accessed on 17 Aug 2010.
[19] `http://www.napster.com`, accessed on 17 Aug 2010.
[20] `http://rfc-gnutella.sourceforge.net/developer/stable/index.html`, accessed on 17 Aug 2010.
[21] `http://rfc-gnutella.sourceforge.net/src/rfc-0_6-draft.html`, accessed on 17 Aug 2010.
[22] `http://g2.trillinux.org/index.php?title=Main_Page`, accessed on 17 Aug 2010.
[23] `www.bearshare.com`, accessed on 17 Aug 2010.

Swapper.NET[24] utilize local directory services that often work with each other for routing queries and merging search results. *Structured-based* P2P networks such as CAN [209] and Chord [252] often use distributed hash tables for searching and retrieving files.

Peer-to-peer search has to address similar problems to federated search; specifically, representing useful contents of peer nodes and local search directories (collection representation), routing queries to relevant nodes or directories (collection selection), and combining search results (result merging). Early P2P networks focused on simple query routing methods such as flooding and simple merging methods based on the frequency of term matching or content-independent features. More recent studies [168, 170, 171] explored full-text representations with content-based query routing and relevance-based results integration. Therefore, improving collection representation, collection selection and result merging in federated search can have a direct impact on the quality of search in P2P networks.

1.2.3 Metasearch Engines

Metasearch engines provide a single search portal for combining the results of multiple search engines [186]. Metasearch engines do not usually retain a document index; they send the query in parallel to multiple search engines, and integrate the returned answers. The architecture details of many metasearch engines such as Dogpile,[25] MetaCrawler [219, 220], AllInOneNews [164], ProFusion [103, 104], Savvysearch [81], iXmetafind [120], Fusion [249], and Inquirus [108, 158] have been published in recent years.

Figure 1.3 shows the answers returned by Metacrawler [220] for the query "federated search". It can be seen that the presented results are merged from different search engines such as Yahoo! and Google, Ask and Bing.

Compared to the centralized search engines, metasearch engines have advantages such as broader coverage of the web and better search scalability [185]. The index and coverage of commercial search engines

[24] `http://www.revolutionarystuff.com/swapper`, accessed on 17 Aug 2010.
[25] `http://www.dogpile.com/dogpile/ws/about?_IceUrl=true`, accessed on 17 Aug 2010.

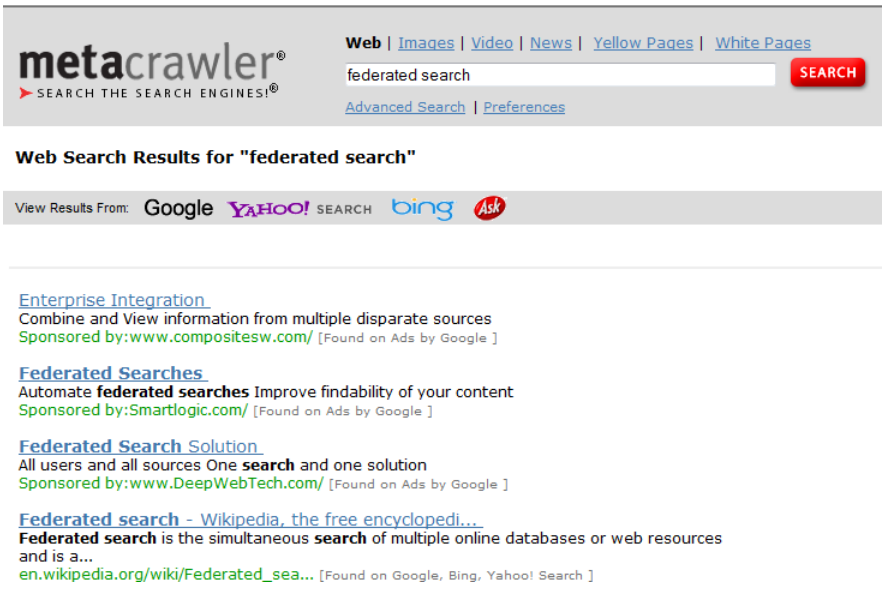

Fig. 1.3 The results of the query "federated search" returned from Metacrawler [220] metasearch engine. It can be seen that the results are merged from different sources such as Google, Yahoo! and Bing search engines.

are substantially different. Many of the pages that are indexed by one search engine may not be indexed by another search engine. Bar-Yossef and Gurevich [22] suggested that the amount of overlap between the indexes of Google and Yahoo! is less than 45%.

1.3 Outline

This paper presents a comprehensive summary of federated search techniques. This section provides a road map for the remaining sections.

In Section 2, we compare the collection representation sets (summaries) in cooperative and uncooperative environments. We also discuss several approaches for improving incomplete summaries, including the previous research on estimating the size of collections from sampled documents. We end this section by describing *wrappers*, the programs used for interacting with the interfaces of hidden-web collections, and summarizing available techniques for evaluating the quality of collection summaries.

In Section 3, we compare different collection selection methods by categorizing the current techniques into two main groups; *lexicon-based*, and *document-surrogates*. The former group mainly consists of techniques that are more suitable for cooperative environments, while the latter group includes collection selection methods based on incomplete sampled documents. We also provide an overview of previous work on query-classification in the context of federated search. In the last section of this section, we discuss the common metrics for evaluating the effectiveness of collection selection methods.

In Section 4, we discuss several federated search merging techniques. We also provide a brief summary of commonly used blending techniques in closely related areas of data fusion and metasearch.

In Section 5, we discuss common datasets used for evaluating the federated search techniques. This is important because relative performance of federated search methods can vary significantly between different testbeds [86, 242].[26]

Finally, in Section 6 we present our conclusions and discuss directions for future work.

[26] We use the term *testbed* to refer to a set of collections that are used together for federated search experiments (collection selection and result merging).

References

[1] F. Abbaci, J. Savoy, and M. Beigbeder, "A methodology for collection selection in heterogeneous contexts," in *Proceedings of the International Conference on Information Technology: Coding and Computing*, pp. 529–535, Washington, DC: IEEE Computer Society, 2002. ISBN 0-7695-1503-1.

[2] D. Aksoy, "Information source selection for resource constrained environments," *SIGMOD Record*, vol. 34, no. 4, pp. 15–20, ISSN 0163-5808, 2005.

[3] J. Allan, J. Aslam, M. Sanderson, C. Zhai, and J. Zobel, eds., *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Boston, MA, 2009. ISBN 978-1-60558-483-6.

[4] J. Allan, M. Connell, and B. Croft, "INQUERY and TREC-9," in *Proceedings of the Ninth Text Retrieval Conference*, (E. Voorhees and D. Harman, eds.), pp. 551–563, Gaithersburg, MD: NIST Special Publication, 2000.

[5] A. Anagnostopoulos, A. Broder, and D. Carmel, "Sampling search-engine results," in *Ellis and Hagino [88]*, pp. 245–256. ISBN 1-59593-046-9.

[6] P. Apers, P. Atzeni, S. Ceri, S. Paraboschi, K. Ramamohanarao, and R. Snodgrass, eds., *Proceedings of the 27th International Conference on Very Large Data Bases*. Roma, Italy: Morgan Kaufmann, 2001. ISBN 1-55860-804-4.

[7] A. Arasu and H. Garcia-Molina, "Extracting structured data from web pages," in *Proceedings ACM SIGMOD International Conference on Management of Data*, (Y. Papakonstantinou and A. Halevy Z. Ives, eds.), pp. 337–348, San Diego, CA, 2003. ISBN 1-58113-634-X.

[8] J. Arguello, J. Callan, and F. Diaz, "Classification-based resource selection," in *Cheung et al. [54]*, pp. 1277–1286. ISBN 978-1-60558-512-3.

[9] J. Arguello, F. Diaz, J. Callan, and J. Crespo, "Sources of evidence for vertical selection," in *Allan et al. [3]*, pp. 315–322. ISBN 978-1-60558-483-6.

[10] H. Ashman and P. Thistlewaite, eds., *Proceedings of the Seventh International Conference on World Wide Web*. Brisbane, Australia: Elsevier, 1998. ISBN 0169-7552.

[11] J. Aslam and M. Montague, "Models for metasearch," in *Croft et al. [72]*, pp. 276–284. ISBN 1-58113-331-6.

[12] J. Aslam, V. Pavlu, and R. Savell, "A unified model for metasearch, pooling, and system evaluation," in *Kraft et al. [152]*, pp. 484–491. ISBN 1-58113-723-0.

[13] T. Avrahami, L. Yau, L. Si, and J. Callan, "The FedLemur: federated search in the real world," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 347–358, ISSN 1532-2882, 2006.

[14] R. Baeza-Yates, *Information retrieval: data structures and algorithms*. Upper Saddle River, NJ: Prentice-Hall, 1992. ISBN 0-13-463837-9.

[15] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA: Addison-Wesley Longman Publishing Co., 1999. ISBN 0-201-39829-X.

[16] P. Bailey, N. Craswell, and D. Hawking, "Engineering a multi-purpose test collection for web retrieval experiments," *Information Processing and Management*, vol. 39, no. 6, pp. 853–871, ISSN 0306-4573, 2003.

[17] M. Baillie, L. Azzopardi, and F. Crestani, "Adaptive query-based sampling of distributed collections," in *Crestani et al. [69]*, pp. 316–328, 2006. ISBN 3-540-45774-7.

[18] M. Baillie, L. Azzopardi, and F. Crestani, "An evaluation of resource description quality measures," in *Proceedings of the ACM symposium on Applied computing*, (H. Haddad, ed.), pp. 1110–1111, Dijon, France, 2006. ISBN 1-59593-108-2.

[19] M. Baillie, L. Azzopardi, and F. Crestani, "Towards better measures: Evaluation of estimated resource description quality for distributed IR," in *Proceedings of the First International Conference on Scalable Information systems*, (X. Jia, ed.), Hong Kong: ACM, 2006. ISBN 1-59593-428-6.

[20] M. Baillie, M. Carman, and F. Crestani, "A topic-based measure of resource description quality for distributed information retrieval," in *Proceedings of the 31st European Conference on Information Retrieval Research,* vol. 5478 of *Lecture Notes in Computer Science*, (M. Boughanem, C. Berrut, J. Mothe, and C. Soulé-Dupuy, eds.), pp. 485–496, Toulouse, France: Springer, 2009.

[21] Z. Bar-Yossef and M. Gurevich, "Efficient search engine measurements," in *Williamson et al. [269]*, pp. 401–410. ISBN 978-1-59593-654-7.

[22] Z. Bar-Yossef and M. Gurevich, "Random sampling from a search engine's index," in *Proceedings of the 15th International Conference on World Wide Web*, (L. Carr, D. Roure, A. Iyengar, C. Goble, and M. Dahlin, eds.), pp. 367–376, Edinburgh, UK: ACM, 2006. ISBN 1-59593-323-9.

[23] L. Barbosa and J. Freire, "Combining classifiers to identify online databases," in *Williamson et al. [269]*. ISBN 978-1-59593-654-7.

[24] C. Baumgarten, "A probabilistic model for distributed information retrieval," in *Belkin et al. [27]*, pp. 258–266. ISBN 0-89791-836-3.

[25] C. Baumgarten, "A probabilistic solution to the selection and fusion problem in distributed information retrieval," in *Gey et al. [106]*, pp. 246–253. ISBN 1-58113-096-1.

[26] N. Belkin, P. Ingwersen, and M. Leong, eds., *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Athens, Greece, 2000. ISBN 1-58113-226-3.

[27] N. J. Belkin, A. D. Narasimhalu, and P. Willett, eds., *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Philadelphia, PL, 1997. ISBN 0-89791-836-3.

[28] M. Bergman, "The deep web: Surfacing hidden value," *Journal of Electronic Publishing*, vol. 7, no. 1, ISSN 1080-2711, 2001.

[29] P. Bernstein, Y. Ioannidis, and R. Ramakrishnan, eds., *Proceedings of the 28th International Conference on Very Large Data Bases*. Hong Kong, China: Morgan Kaufmann, 2002.

[30] Y. Bernstein, M. Shokouhi, and J. Zobel, "Compact features for detection of near-duplicates in distributed retrieval," in *Crestani et al. [69]*, pp. 110–121. ISBN 3-540-45774-7.

[31] Y. Bernstein and J. Zobel, "A scalable system for identifying co-derivative documents," in *Proceedings of the 11th International String Processing and Information Retrieval Conference,* vol. 3246 of *Lecture Notes in Computer Science*, (A. Apostolico and M. Melucci, eds.), pp. 55–67, Padova, Italy: Springer, 2004. ISBN 3-540-23210-9.

[32] S. Berretti, J. Callan, H. Nottelmann, X. M. Shou, and S. Wu, "MIND: resource selection and data fusion in multimedia distributed digital libraries," in *Clarke et al. [56]*, pp. 465–465. ISBN 1-58113-646-3.

[33] K. Bharat and A. Broder, "A technique for measuring the relative size and overlap of public web search engines," in *Ashman and Thistlewaite [10]*, pp. 379–388, 1998. ISBN 0169-7552.

[34] K. Bharat and A. Broder, "A technique for measuring the relative size and overlap of public web search engines," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 379–388, ISSN 0169-7552, 1998.

[35] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, ISSN 1533-7928, 2003.

[36] S. Brin, J. Davis, and H. García-Molina, "Copy detection mechanisms for digital documents," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, (M. Carey and D. Schneider, eds.), pp. 398–409, San Jose, CA, 1995. ISBN 0-89791-731-6.

[37] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Ashman and Thistlewaite [10]*, pp. 107–117. ISBN 0169-7552.

[38] A. Broder, M. Fontura, V. Josifovski, R. Kumar, R. Motwani, S. Nabar, R. Panigrahy, A. Tomkins, and Y. Xu, "Estimating corpus size via queries," in *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*, (P. Yu, V. Tsotras, E. Fox, and B. Liu, eds.), pp. 594–603, Arlington, VA, 2006. ISBN 1-59593-433-2.

[39] A. Broder, S. Glassman, M. Manasse, and G. Zweig, "Syntactic clustering of the web," *Computer Networks and ISDN System*, vol. 29, no. 8–13, pp. 1157–1166, ISSN 0169-7552, 1997.

[40] D. Buttler, L. Liu, and C. Pu, "A fully automated object extraction system for the World Wide Web," in *Shen et al. [224]*, pp. 361–370. ISBN 1-58113-348-0.

[41] J. Callan, "Distributed information retrieval," in *Advances in information retrieval, Chapter 5,* vol. 7 of *The Information Retrieval Series*, (B. Croft, ed.), pp. 127–150, Kluwer Academic Publishers, 2000. ISBN 978-0-7923-7812-9.

[42] J. Callan and M. Connell, "Query-based sampling of text databases," *ACM Transactions on Information Systems*, vol. 19, no. 2, pp. 97–130, ISSN 1046-8188, 2001.

[43] J. Callan, M. Connell, and A. Du, "Automatic discovery of language models for text databases," in *Proceedings ACM SIGMOD International Conference on Management of Data*, (A. Delis, C. Faloutsos, and S. Ghandeharizadeh, eds.), pp. 479–490, Philadelphia, PA, 1999. ISBN 1-58113-084-8.

[44] J. Callan, F. Crestani, and M. Sanderson, eds., *Distributed Multimedia Information Retrieval, SIGIR 2003 Workshop on Distributed Information Retrieval, Revised Selected and Invited Papers,* volume 2924 of *Lecture Notes in Computer Science*. Toronto, Canada: Springer, 2004. ISBN 3-540-20875-5.

[45] J. Callan, B. Croft, and S. Harding, "The INQUERY retrieval system," in *Proceedings of Third International Conference on Database and Expert Systems Applications*, (A. Tjoa and I. Ramos, eds.), pp. 78–83, Valencia, Spain, 1992. ISBN 3-211-82400-6.

[46] J. Callan, Z. Lu, and B. Croft, "Searching distributed collections with inference networks," in *Fox et al. [91]*, pp. 21–28. ISBN 0-89791-714-6.

[47] J. Callan, A. Powell, J. French, and M. Connell, "The effects of query-based sampling on automatic database selection algorithms," Technical Report, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2000.

[48] A. Calvé and J. Savoy, "Database merging strategy based on logistic regression," *Information Processing and Management*, vol. 36, no. 3, pp. 341–359, ISSN 0306-4573, 2000.

[49] M. Carman and F. Crestani, "Towards personalized distributed information retrieval," in *Myaeng et al. [192]*, pp. 719–720. ISBN 978-1-60558-164-4.

[50] J. Caverlee, L. Liu, and J. Bae, "Distributed query sampling: A quality-conscious approach," in *Efthimiadis et al. [87]*, pp. 340–347. ISBN 1-59593-369-7.

[51] S. Cetinta, L. Si, and H. Yuan, "Learning from past queries for resource selection," in *Cheung et al. [54]*, pp. 1867–1870. ISBN 978-1-60558-512-3.

[52] A. Chakravarthy and K. Haase, "NetSerf: using semantic knowledge to find internet information archives," in *Fox et al. [91]*, pp. 4–11. ISBN 0-89791-714-6.

[53] C. H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey of web information extraction systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1411–1428, ISSN 1041-4347, 2006.

[54] D. Cheung, I. Song, W. Chu, X. Hu, and J. Lin, eds., *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*. Hong Kong, China, 2009. ISBN 978-1-60558-512-3.

[55] J. Cho and H. Garcia-Molina, "Effective page refresh policies for web crawlers," *ACM Transactions on Database Systems*, vol. 28, no. 4, pp. 390–426, ISSN 0362-5915, 2003.

[56] C. Clarke, G. Cormack, J. Callan, D. Hawking, and A. Smeaton, eds., *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Toronto, Canada, 2003. ISBN 1-58113-646-3.

[57] C. Clarke, N. Craswell, and I. Soboroff, "The TREC terabyte retrieval track," *SIGIR Forum*, vol. 39, no. 1, pp. 31–47, ISSN 0163-5840, 2005.

[58] W. Cohen, "Learning trees and rules with set-valued features," in *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI)*, pp. 709–716, Portland, OR, 1996. ISBN 0-262-51091-X.

[59] J. Conrad and J. Claussen, "Early user — system interaction for database selection in massive domain-specific online environments," *ACM Transactions on Information Systems*, vol. 21, no. 1, pp. 94–131, ISSN 1046-8188, 2003.

[60] J. Conrad, X. Guo, P. Jackson, and M. Meziou, "Database selection using actual physical and acquired logical collection resources in a massive domain-specific operational environment," in *Bernstein et al. [29]*, pp. 71–82.

[61] J. Conrad, C. Yang, and J. Claussen, "Effective collection metasearch in a hierarchical environment: global vs. localized retrieval performance," in *Järvelin et al. [142]*. ISBN 1-58113-561-0.

[62] J. Cope, N. Craswell, and D. Hawking, "Automated discovery of search interfaces on the web," in *Proceedings of the 14th Australasian database conference,* vol. 17, p. 189, Australian Computer Society, Inc., 2003.

[63] N. Craswell, "Methods for distributed information retrieval," PhD thesis, Australian National University, 2000.

[64] N. Craswell, P. Bailey, and D. Hawking, "Server selection on the World Wide Web," in *Proceedings of the Fifth ACM Conference on Digital Libraries*, pp. 37–46, San Antonio, TX, 2000. ISBN 1-58113-231-X.

[65] N. Craswell and D. Hawking, "Overview of the TREC-2002 web track," in *Proceedings of the 11th Text REtrieval Conference*, (E. Voorhees, ed.), pp. 86–95, Gaithersburg, MD: NIST Special Publication, 2002.

[66] N. Craswell, D. Hawking, and S. Robertson, "Effective site finding using link anchor information," in *Croft et al. [72]*, pp. 250–257. ISBN 1-58113-331-6.

[67] N. Craswell, D. Hawking, and P. Thistlewaite, "Merging results from isolated search engines," in *Proceedings of the 10th Australasian Database Conference*, pp. 189–200, Auckland, New Zealand: Springer, 1999. ISBN 981-4021-55-5.

[68] V. Crescenzi, G. Mecca, and P. Merialdo, "Roadrunner: Towards automatic data extraction from large web sites," in *Apers et al. [6]*, pp. 109–118. ISBN 1-55860-804-4.

[69] F. Crestani, P. Ferragina, and M. Sanderson, eds., *Proceedings of the 13th International String Processing and Information Retrieval Conference,* vol. 4209 of *Lecture Notes in Computer Science*. Glasgow, UK: Springer, 2006. ISBN 3-540-45774-7.

[70] F. Crestani, S. Marchand-Maillet, H. Chen, E. Efthimiadis, and J. Savoy, eds., *Proceedings of the 33rd Annual International ACM SIGIR Conference*

on Research and Development in Information Retrieval. Geneva, Switzerland, 2010. ISBN 978-1-4503-0153-4.

[71] B. Croft, "Combining approaches to information retrieval," in *Advances in Information Retrieval, Chapter 1,* volume 7 of *The Information Retrieval Series*, (B. Croft, ed.), pp. 1–36, Kluwer Academic Publishers, 2000. ISBN 978-0-7923-7812-9.

[72] B. Croft, D. Harper, D. H. Kraft, and J. Zobel, eds., *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New Orleans, LA, 2001. ISBN 1-58113-331-6.

[73] B. Croft, A. Moffat, K. Rijsbergen, R. Wilkinson, and J. Zobel, eds., *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Melbourne, Australia, 1998. ISBN 1-58113-015-5.

[74] O. de Kretser, A. Moffat, T. Shimmin, and J. Zobel, "Methodologies for distributed information retrieval," in *Proceedings of the Eighteenth International Conference on Distributed Computing Systems*, (M. Papazoglou, M. Takizawa, B. Kramer, and S. Chanson, eds.), pp. 66–73, Amsterdam, The Netherlands, 1998. ISBN 0-8186-8292-2.

[75] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Sciences*, vol. 41, no. 6, pp. 391–407, 1990.

[76] M. DeGroot, *Optimal Statistical Decisions (Wiley Classics Library)*. Wiley interscience, 2004. ISBN 978-0-471-72614-2.

[77] F. Diaz, "Integration of news content into web results," in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pp. 182–191, Barcelona, Spain: ACM, 2009. ISBN 978-1-60558-390-7.

[78] F. Diaz and J. Arguello, "Adaptation of offline vertical selection predictions in the presence of user feedback," in *Allan et al. [3]*, pp. 323–330. ISBN 978-1-60558-483-6.

[79] F. Diaz and J. Arguello, "Adaptation of offline vertical selection predictions in the presence of user feedback," in *Crestani et al. [70]*, pp. 323–330. ISBN 978-1-4503-0153-4.

[80] F. Diaz and D. Metzler, "Improving the estimation of relevance models using large external corpora," in *Efthimiadis et al. [87]*, pp. 154–161. ISBN 1-59593-369-7.

[81] D. Dreilinger and A. Howe, "Experiences with selecting search engines using metasearch," *ACM Transaction on Information Systems*, vol. 15, no. 3, pp. 195–222, ISSN 1046-8188, 1997.

[82] D. D'Souza, "Document retrieval in managed document collections," PhD thesis, RMIT University, Melbourne, Australia, 2005.

[83] D. D'Souza and J. Thom, "Collection selection using n-term indexing," in *Proceedings of the Second International Symposium on Cooperative Database Systems for Advanced Applications (CODAS'99)*, (Y. Zhang, M. Rusinkiewicz, and Y. Kambayashi, eds.), pp. 52–63, Wollongong, NSW, Australia: Springer, 1999. ISBN 9814021644.

[84] D. D'Souza, J. Thom, and J. Zobel, "A comparison of techniques for selecting text collections," in *Proceedings of the Australasian Database Conference*,

p. 28, Canberra, Australia: IEEE Computer Society, 2000. ISBN 0-7695-0528-7.

[85] D. D'Souza, J. Thom, and J. Zobel, "Collection selection for managed distributed document databases," *Information Processing and Management*, vol. 40, no. 3, pp. 527–546, ISSN 0306-4573, 2004.

[86] D. D'Souza, J. Zobel, and J. Thom, "Is CORI effective for collection selection? An exploration of parameters, queries, and data," in *Proceedings of the Australaian Document Computing Symposium*, (P. Bruza, A. Moffat, and A. Turpin, eds.), pp. 41–46, Melbourne, Australia, 2004. Melbourne, Australia. ISBN 0-9757172-0-0.

[87] E. Efthimiadis, S. Dumais, D. Hawking, and K. Järvelin, eds., *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Seattle, WA, 2006. ISBN 1-59593-369-7.

[88] A. Ellis and T. Hagino, eds., *Proceedings of the 14th International Conference on World Wide Web*. Chiba, Japan: ACM, 2005. ISBN 1-59593-046-9.

[89] J. Elsas, J. Arguello, J. Callan, and J. Carbonell, "Retrieval and feedback models for blog feed search," in *Myaeng et al. [192]*, pp. 347–354. ISBN 978-1-60558-164-4.

[90] D. Fetterly, M. Manasse, and M. Najork, "On the evolution of clusters of near-duplicate web pages," in *Proceedings of the First Conference on Latin American Web Congress*, p. 37, Washington, DC: IEEE Computer Society, 2003. ISBN 0-7695-2058-8.

[91] E. Fox, P. Ingwersen, and R. Fidel, eds., *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Seattle, WA, 1995. ISBN 0-89791-714-6.

[92] E. Fox and J. Shaw, "Combination of multiple searches," in *Proceedings of the Second Text REtrieval Conference*, (D. Harman, ed.), pp. 243–252, Gaithersburg, MD, 1993. NIST Special Publication.

[93] E. Fox and J. Shaw, "Combination of multiple searches," in *Proceedings of the Third Text REtrieval Conference*, (D. Harman, ed.), pp. 105–108, Gaithersburg, MD, 1994. NIST Special Publication.

[94] J. French and A. Powell, "Metrics for evaluating database selection techniques," *World Wide Web*, vol. 3, no. 3, pp. 153–163, ISSN 1386-145X, 2000.

[95] J. French, A. Powell, J. Callan, C. Viles, T. Emmitt, K. Prey, and Y. Mou, "Comparing the performance of database selection algorithms," in *Gey et al. [106]*, pp. 238–245. ISBN 1-58113-096-1.

[96] J. French, A. Powell, F. Gey, and N. Perelman, "Exploiting a controlled vocabulary to improve collection selection and retrieval effectiveness," in *Paques et al. [200]*, pp. 199–206. ISBN 1-58113-436-3.

[97] J. French, A. Powell, C. Viles, T. Emmitt, and K. Prey, "Evaluating database selection techniques: A testbed and experiment," in *Croft et al. [73]*, pp. 121–129. ISBN 1-58113-015-5.

[98] N. Fuhr, "Optimum database selection in networked IR," in *Proceedings of the SIGIR'96 Workshop on Networked Information Retrieval (NIR'96)*, (J. Callan and N. Fuhr, eds.), Zurich, Switzerland, 1996.

[99] N. Fuhr, "A decision-theoretic approach to database selection in networked IR," *ACM Transactions on Information Systems*, vol. 17, no. 3, pp. 229–249, ISSN 1046-8188, 1999.

[100] N. Fuhr, "Resource discovery in distributed digital libraries," in *Proceedings of Digital Libraries Advanced Methods and Technologies, Digital Collections*, pp. 35–45, Petersburg, Russia, 1999.

[101] S. Garcia, H. Williams, and A. Cannane, "Access-ordered indexes," in *Proceedings of the 27th Australasian Computer Science Conference*, (V. Estivill-Castro, ed.), pp. 7–14, Darlinghurst, Australia: Australian Computer Society, 2004. ISBN 1-920682-05-8.

[102] S. Gauch, ed., *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*. Kansas, MO, 1999. ISBN 1-58113-1461.

[103] S. Gauch and G. Wang, "Information fusion with ProFusion," in *Proceedings of the First World Conference of the Web Society*, pp. 174–179, San Francisco, CA, 1996.

[104] S. Gauch, G. Wang, and M. Gomez, "ProFusion: Intelligent fusion from multiple distributed search engines," *Journal of Universal Computer Science*, vol. 2, no. 9, pp. 637–649, ISSN 0948-695X, 1996.

[105] S. Gauch, G. Wang, and M. Gomez, "ProFusion: Intelligent fusion from multiple, distributed search engines," *Journal of Universal Computer Science*, vol. 2, no. 9, pp. 637–649, ISSN 1041-4347, 1996.

[106] F. Gey, M. Hearst, and R. Tong, eds., *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Berkeley, CA, 1999. ISBN 1-58113-096-1.

[107] E. Glover and S. Lawrence, *Selective Retrieval Metasearch Engine (United States Patent 2002/0165860 a1)*. 2001.

[108] E. Glover, S. Lawrence, W. Birmingham, and C. Giles, "Architecture of a metasearch engine that supports user information needs," in *Gauch [102]*, pp. 210–216, 1999. ISBN 1-58113-1461.

[109] J. Goldberg, "CDM: An approach to learning in text categorization," in *Proceedings of the Seventh International Tools with Artificial Intelligence*, pp. 258–265, Herndon, VA: IEEE Computer Society, 1995. ISBN 0-8186-7312-5.

[110] L. Gravano, "Querying multiple document collections across the internet," PhD thesis, Stanford University, 1997.

[111] L. Gravano, C. Chang, H. García-Molina, and A. Paepcke, "STARTS: Stanford proposal for Internet meta-searching," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, (J. Peckham, ed.), pp. 207–218, Tucson, AZ, 1997. ISBN 0-89791-911-4.

[112] L. Gravano and H. García-Molina, "Generalizing GlOSS to vector-space databases and broker hierarchies," in *Proceedings of the 21st International Conference on Very Large Data Bases*, (U. Dayal, P. Gray, and S. Nishio, eds.), pp. 78–89, Zurich, Switzerland: Morgan Kaufmann, 1995. ISBN 1-55860-379-4.

[113] L. Gravano, H. García-Molina, and A. Tomasic, "The effectiveness of GlOSS for the text database discovery problem," in *Proceedings of the ACM*

SIGMOD International Conference on Management of Data, (R. Snodgrass and M. Winslett, eds.), pp. 126–137, Minneapolis, MN, 1994. ISBN 0-89791-639-5.

[114] L. Gravano, H. García-Molina, and A. Tomasic, "Precision and recall of GlOSS estimators for database discovery," in *Proceedings of the Third International Conference on Parallel and Distributed Information Systems*, pp. 103–106, Austin, TX: IEEE Computer Society, 1994. ISBN 0-8186-6400-2.

[115] L. Gravano, H. García-Molina, and A. Tomasic, "GlOSS: text-source discovery over the Internet," *ACM Transactions on Database Systems*, vol. 24, no. 2, pp. 229–264, ISSN 0362-5915, 1999.

[116] L. Gravano, P. Ipeirotis, and M. Sahami, "Qprober: A system for automatic classification of hidden web databases," *ACM Transactions on Information Systems*, vol. 21, no. 1, pp. 1–41, ISSN 1046-8188, 2003.

[117] N. Green, P. Ipeirotis, and L. Gravano, "SDLIP + STARTS = SDARTS a protocol and toolkit for metasearching," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, (E. Fox and C. Borgman, eds.), pp. 207–214, Roanoke, VA: ACM, 2001. ISBN 1-58113-345-6.

[118] J. Gross, *Linear Regression*. Springer, 2003. ISBN 3540401784.

[119] A. Gulli and A. Signorini, "The indexable web is more than 11.5 billion pages," in *Ellis and Hagino [88]*, pp. 902–903. ISBN 1-59593-046-9.

[120] E. Han, G. Karypis, D. Mewhort, and K. Hatchard, "Intelligent metasearch engine for knowledge management," in *Kraft et al. [152]*, pp. 492–495, 2003. ISBN 1-58113-723-0.

[121] D. Harman, "Overview of the Third Text REtrieval Conference (TREC-3)," in *Proceedings of the Third Text REtrieval Conference*, (D. Harman, ed.), pp. 1–19, Gaithersburg, MD, 1994. NIST Special Publication.

[122] D. Harman, "Overview of the fourth Text REtrieval Conference (TREC-4)," in *Proceedings of the Fourth Text REtrieval Conference*, (D. Harman, ed.), pp. 1–24, Gaithersburg, MD, 1995. NIST Special Publication.

[123] D. Hawking and P. Thistlewaite, "Overview of TREC-6 very large collection track," in *Proceedings of the Seventh Text REtrieval Conference*, (E. Voorhees and D. Harman, eds.), pp. 93–106, Gaithersburg, MD, 1997. NIST Special Publication.

[124] D. Hawking and P. Thistlewaite, "Methods for information server selection," *ACM Transactions on Information Systems*, vol. 17, no. 1, pp. 40–76, ISSN 1046-8188, 1999.

[125] D. Hawking and P. Thomas, "Server selection methods in hybrid portal search," in *Marchionini et al. [181]*, pp. 75–82, 2005. ISBN 1-59593-034-5.

[126] D. Hawking, E. Voorhees, N. Craswell, and P. Bailey, "Overview of the TREC-8 web track," in *Proceedings of the Eight Text REtrieval Conference*, (E. Voorhees and D. Harman, eds.), pp. 131–150, Gaithersburg, MD, 2000. NIST Special Publication.

[127] Y. Hedley, M. Younas, A. James, and M. Sanderson, "Information extraction from template-generated hidden web documents," in *Proceedings of the IADIS International Conference WWW/Internet*, (P. Isaías, N. Karmakar,

L. Rodrigues, and P. Barbosa, eds.), pp. 627–634, Madrid, Spain, 2004. ISBN 972-99353-0-0.

[128] Y. Hedley, M. Younas, A. James, and M. Sanderson, "Query-related data extraction of hidden web documents," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, eds.), pp. 558–559, Sheffield, UK, 2004. ISBN 1-58113-881-4.

[129] Y. Hedley, M. Younas, A. James, and M. Sanderson, "A two-phase sampling technique for information extraction from hidden web databases," in *Proceedings of the Sixth Annual ACM International Workshop on Web Information and Data Management*, (A. Laender and D. Lee, eds.), pp. 1–8, Washington DC, 2004. ISBN 1-58113-978-0.

[130] Y. Hedley, M. Younas, A. James, and M. Sanderson, "A two-phase sampling technique to improve the accuracy of text similarities in the categorisation of hidden web databases," in *Proceedings of the Fifth International Conference on Web Information Systems Engineering,* vol. 3306 of *Lecture Notes in Computer Science*, (X. Zhou, S. Su, M. Papazoglou, M. Orlowska, and K. Jeffery, eds.), pp. 516–527, Brisbane, Australia: Springer, 2004. ISBN 3-540-23894-8.

[131] M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork, "On near-uniform url sampling," in *Herman and Vezza [132]*, pp. 295–308. ISBN 1-930792-01-8.

[132] I. Herman and A. Vezza, eds., *Proceedings of the Ninth International Conference on World Wide Web*. Amsterdam, The Netherlands: Elsevier, 2000. ISBN 1-930792-01-8.

[133] T. Hernandez and S. Kambhampati, "Improving text collection selection with coverage and overlap statistics," in *Ellis and Hagino [88]*, pp. 1128–1129. ISBN 1-59593-046-9.

[134] D. Hong, L. Si, P. Bracke, M. Witt Michael, and T. Juchcinski, "A joint probabilistic classification model for resource selection," in *Crestani et al. [70]*, pp. 98–105. ISBN 978-1-4503-0153-4.

[135] D. Hosmer and S. Lemeshow, *Applied Logistic Regression*. New York, NY: John Wiley & Sons, 1989. ISBN 0-471-35632-8.

[136] P. Ipeirotis, "Classifying and searching hidden-web text databases," PhD thesis, Columbia University, 2004.

[137] P. Ipeirotis and L. Gravano, "Distributed search over the hidden web: Hierarchical database sampling and selection," in *Bernstein et al. [29]*, pp. 394–405.

[138] P. Ipeirotis and L. Gravano, "When one sample is not enough: improving text database selection using shrinkage," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, (G. Weikum, A. König, and S. Deßloch, eds.), pp. 767–778, Paris, France, 2004. ISBN 1-58113-859-8.

[139] P. Ipeirotis and L. Gravano, "Classification-aware hidden-web text database selection," *ACM Transactions on Information Systems*, vol. 26, no. 2, pp. 1–66, ISSN 1046-8188, 2008.

[140] P. Ipeirotis, A. Ntoulas, J. Cho, and L. Gravano, "Modeling and managing content changes in text databases," in *Proceedings of the 21st International Conference on Data Engineering*, pp. 606–617, Tokyo, Japan: IEEE, 2005. ISBN 0-7695-2285-8.

[141] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall, 1988. ISBN 0-13-022278-X.

[142] K. Järvelin, M. Beaulieu, R. Baeza-Yates, and S. Myaeng, eds., *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Tampere, Finland, 2002. ISBN 1-58113-561-0.

[143] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in *Belkin et al. [26]*, pp. 41–48. ISBN 1-58113-226-3.

[144] K. Kalpakis, N. Goharian, and D. Grossman, eds., *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*. McLean, VA, 2002. ISBN 1-58113-492-4.

[145] S. Karnatapu, K. Ramachandran, Z. Wu, B. Shah, V. Raghavan, and R. Benton, "Estimating size of search engines in an uncooperative environment," in *Proceedings of the Second International Workshop on Web-based Support Systems*, (J. Yao, V. Raghavan, and G. Wang, eds.), pp. 81–87, Beijing, China: Saint Mary's University, Canada, 2004. ISBN 0-9734039-6-9.

[146] J. Kim and B. Croft, "Ranking using multiple document types in desktop search," in *Crestani et al. [70]*, pp. 50–57. ISBN 978-1-4503-0153-4.

[147] J. King, P. Bruza, and R. Nayak, "Preliminary investigations into ontology-based collection selection," in *Proceedings of the 11th Australasian Document Computing Symposium*, (P. Bruza, A. Spink, and R. Wilkinson, eds.), pp. 33–40, Brisbane, Australia, 2006. ISBN 1-74107-140-2.

[148] T. Kirsch, 2003. Document retrieval over networks wherein ranking and relevance scores are computed at the client for multiple database documents (United States Patent 5,659,732).

[149] A. König, M. Gamon, and Q. Wu, "Click-through prediction for news queries," in *Allan et al. [3]*, pp. 347–354. ISBN 978-1-60558-483-6.

[150] M. Koster, "ALIWEB, Archie-like indexing in the web," *Computer Networks and ISDN Systems*, vol. 27, no. 2, pp. 175–182, ISSN 1389-1286, 1994.

[151] W. Kraaij, A. de Vries, C. Clarke, N. Fuhr, and N. Kando, eds., *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Amsterdam, The Netherlands, 2007. ISBN 978-1-59593-597-7.

[152] D. Kraft, O. Frieder, J. Hammer, S. Qureshi, and L. Seligman, eds., *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*. New Orleans, LA, 2003. ISBN 1-58113-723-0.

[153] S. Kullback, *Information Theory and Statistics*. New York, NY: John Wiley & Sons, 1959. ISBN 0486696847.

[154] J. Lafferty and C. Zhai, "Document language models, query models, and risk minimization for information retrieval," in *Croft et al. [72]*, pp. 111–119. ISBN 1-58113-331-6.

[155] L. Larkey, M. Connell, and J. Callan, "Collection selection and results merging with topically organized U.S. patents and TREC data," in *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*, (A. Agah, J. Callan, E. Rundensteiner, and S. Gauch, eds.), pp. 282–289, McLean, VA, 2000. ISBN 1-58113-320-0.

[156] R. Larson, "A logistic regression approach to distributed IR," in *Järvelin et al. [142]*, pp. 399–400. ISBN 1-58113-561-0.

[157] R. Larson, "Distributed IR for digital libraries," in *Research and Advanced Technology for Digital Libraries, Seventh European Conference,* vol. 2769 of *Lecture Notes in Computer Science*, (T. Koch and I. Sølvberg, eds.), pp. 487–498, Trondheim, Norway: Springer, 2003. ISBN 3-540-40726-X.

[158] S. Lawrence and C. Giles, "Inquirus, the NECi meta search engine," in *Ashman and Thistlewaite [10]*, pp. 95–105. ISBN 0169-7552.

[159] J. Lee, "Analyses of multiple evidence combination," in *Belkin et al. [27]*, pp. 267–276. ISBN 0-89791-836-3.

[160] D. Lillis, F. Toolan, R. Collier, and J. Dunnion, "ProbFuse: a probabilistic approach to data fusion," in *Efthimiadis et al. [87]*, pp. 139–146. ISBN 1-59593-369-7.

[161] D. Lillis, F. Toolan, R. Collier, and J. Dunnion, "Extending probabilistic data fusion using sliding windows," in *Proceedings of the 30th European Conference on Information Retrieval Research,* volume 4956 of *Lecture Notes in Computer Science*, (C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. White, eds.), pp. 358–369, Glasgow, UK: Springer, 2008.

[162] K. Lin and H. Chen, "Automatic information discovery from the invisible web," in *Proceedings of the International Conference on Information Technology: Coding and Computing*, pp. 332–337, Washington, DC: IEEE Computer Society, 2002. ISBN 0-7695-1503-1.

[163] B. Liu, R. Grossman, and Y. Zhai, "Mining data records in web pages," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (L. Getoor, T. E. Senator, P. Domingos, and C. Faloutsos, eds.), Washington, DC, USA, 2003. ISBN 1-58113-737-0.

[164] K. Liu, W. Meng, J. Qiu, C. Yu, V. Raghavan, Z. Wu, Y. Lu, H. He, and H. Zhao, "Allinonenews: development and evaluation of a large-scale news metasearch engine," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, (C. Y. Chan, B. C. Ooi, and A. Zhou, eds.), pp. 1017–1028, Beijing, China, 2007. ISBN 978-1-59593-686-8.

[165] K. Liu, C. Yu, and W. Meng, "Discovering the representative of a search engine," in *Paques et al. [200]*, pp. 652–654. ISBN 1-58113-436-3.

[166] W. Liu, X. Meng, and W. Meng, "Vide: A vision-based approach for deep web data extraction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 3, ISSN 1041-4347, 2010.

[167] J. Lu, "Efficient estimation of the size of text deep web data source," in *Shanahan et al. [223]*, pp. 1485–1486. ISBN 978-1-59593-991-3.

[168] J. Lu, "Full-text federated search in peer-to-peer networks," PhD thesis, Carnegie Mellon University, 2007.

[169] J. Lu and J. Callan, "Pruning long documents for distributed information retrieval," in *Kalpakis et al. [144]*, pp. 332–339. ISBN 1-58113-492-4.

[170] J. Lu and J. Callan, "User modeling for full-text federated search in peer-to-peer networks," in *Efthimiadis et al. [87]*, pp. 332–339. ISBN 1-59593-369-7.

[171] J. Lu and J. Callan, "Content-based retrieval in hybrid peer-to-peer networks," in *Kraft et al. [152]*, pp. 199–206, 2003. ISBN 1-58113-723-0.

[172] J. Lu and J. Callan, "Reducing storage costs for federated search of text databases," in *Proceedings of the 2003 Annual national Conference on Digital Government Research*, pp. 1–6, Boston, MA: Digital Government Research Center, 2003.

[173] J. Lu and J. Callan, "Federated search of text-based digital libraries in hierarchical peer-to-peer networks," in *Proceedings of the 27th European Conference on IR Research*, (D. Losada and J. Fernández-Luna, eds.), pp. 52–66, Santiago de Compostela, Spain: Springer, 2005. ISBN 3-540-25295-9.

[174] J. Lu and D. Li, "Estimating deep web data source size by capturerecapture method," *Information Retrieval*, page to appear, ISSN 1386-4564, 2009.

[175] J. Lu, Y. Wang, J. Liang, J. Chen, and J. Liu, "An approach to deep web crawling by sampling," *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 718–724, 2008.

[176] Y. Lu, W. Meng, L. Shu, C. Yu, and K. Liu, "Evaluation of result merging strategies for metasearch engines," in *Proceedings of the Sixth International Conference on Web Information Systems Engineering,* vol. 3806 of *Lecture Notes in Computer Science*, (A. Ngu, M. Kitsuregawa, E. Neuhold, J. Chung, and Q. Sheng, eds.), pp. 53–66, New York, NY: Springer, 2005. ISBN 3-540-30017-1.

[177] J. Madhavan, S. Jeffery, S. Cohen, X. Dong, D. Ko, C. Yu, and A. Halevy, "Web-scale data integration: You can only afford to pay as you go," in *Proceedings of Conference on Innovative Data Systems Research*, pp. 342–350, 2007.

[178] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy, "Google's Deep Web crawl," *Proceedings of VLDB*, vol. 1, no. 2, pp. 1241–1252, 2008.

[179] U. Manber, "Finding similar files in a large file system," in *Proceedings of the USENIX Winter Technical Conference*, pp. 1–10, San Fransisco, CA, 1994. ISBN 1-880446-58-8.

[180] U. Manber and P. Bigot, "The search broker," in *USENIX Symposium on Internet Technologies and Systems*, Monterey, CA, 1997.

[181] G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, eds., *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Salvador, Brazil, 2005. ISBN 1-59593-034-5.

[182] Z. Mazur, "On a model of distributed information retrieval systems based on thesauri," *Information Processing and Management*, vol. 20, no. 4, pp. 499–505, ISSN 0306-4573, 1984.

[183] A. McCallum, R. Rosenfeld, T. Mitchelland, and A. Ng, "Improving text classification by shrinkage in a hierarchy of classes," in *Proceedings of the 15th International Conference on Machine Learning*, pp. 359–367, San Francisco, CA: Morgan Kaufmann Publishers Inc., 1998. ISBN 1-55860-556-8.

[184] W. Meng, Z. Wu, C. Yu, and Z. Li, "A highly scalable and effective method for metasearch," *ACM Transactions on Information Systems*, vol. 19, no. 3, pp. 310–335, ISSN 1046-8188, 2001.

[185] W. Meng, C. Yu, and K. Liu, "Building efficient and effective metasearch engines," *ACM Computing Surveys*, vol. 34, no. 1, pp. 48–89, ISSN 0360-0300, 2002.

[186] W. Y. Meng and C. Yu, *Advanced Metasearch Engine Technology*. Morgan & Claypool Publishers, 2010. ISBN 1608451925.

[187] D. Metzler and B. Croft, "Latent concept expansion using markov random fields," in *Kraaij et al. [151]*, pp. 311–318. ISBN 978-1-59593-597-7.

[188] A. Moffat and J. Zobel, "Information retrieval systems for large document collections," in *Proceedings of the Third Text REtrieval Conference*, (D. Harman, ed.), pp. 85–94, Gaithersburg, MD, 1994. NIST Special Publication.

[189] G. Monroe, J. French, and A. Powell, "Obtaining language models of web collections using query-based sampling techniques," in *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 3*, pp. 1241–1247, Honolulu, HI: IEEE Computer Society, 2002. ISBN 0-7695-1435-9.

[190] G. Monroe, D. Mikesell, and J. French, "Determining stopping criteria in the generation of web-derived language models," Technical report, University of Viginia, 2000.

[191] V. Murdock and M. Lalmas, "Workshop on aggregated search," *SIGIR Forum*, vol. 42, no. 2, pp. 80–83, ISSN 0163-5840, 2008.

[192] W. Myaeng, D. Oard, F. Sebastiani, T. Chua, and M. Leong, eds., *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Singapore, 2008. ISBN 978-1-60558-164-4.

[193] K. Ng, "An investigation of the conditions for effective data fusion in information retrieval," PhD thesis, Rutgers University, 1998.

[194] H. Nottelmann and N. Fuhr, "Decision-theoretic resource selection for different data types in MIND," in *Callan et al. [44]*, pp. 43–57. ISBN 3-540-20875-5.

[195] H. Nottelmann and N. Fuhr, "Evaluating different methods of estimating retrieval quality for resource selection," in *Clarke et al. [56]*, pp. 290–297. ISBN 1-58113-646-3.

[196] H. Nottelmann and N. Fuhr, "The MIND architecture for heterogeneous multimedia federated digital libraries," in *Callan et al. [44]*, pp. 112–125. ISBN 3-540-20875-5.

[197] H. Nottelmann and N. Fuhr, "Combining CORI and the decision-theoretic approach for advanced resource selection," in *Proceedings of the 26th European Conference on IR Research,* vol. 2997 of *Lecture Notes in Computer Science*, (S. McDonald and J. Tait, eds.), pp. 138–153, Sunderland, UK: Springer, 2004. ISBN 3-540-21382-1.

[198] P. Ogilvie and J. Callan, "The effectiveness of query expansion for distributed information retrieval," in *Paques et al. [200]*, pp. 183–190. ISBN 1-58113-436-3.

[199] B. Oztekin, G. Karypis, and V. Kumar, "Expert agreement and content based reranking in a meta search environment using mearf," in *Proceedings of the 11th International Conference on World Wide Web*, (D. Lassner, D. Roure, and A. Iyengar, eds.), pp. 333–344, Honolulu, HI: ACM, 2002. ISBN 1-58113-449-5.

[200] H. Paques, L. Liu, and D. Grossman, eds., *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*. Atlanta, GA, 2001. ISBN 1-58113-436-3.

[201] J. Ponte and B. Croft, "A language modeling approach to information retrieval," in *Croft et al. [73]*, pp. 275–281. ISBN 1-58113-015-5.

[202] A. Powell, "Database selection in distributed information retrieval: A study of multi-collection information retrieval," PhD thesis, University of Virginia, 2001.

[203] A. Powell and J. French, "Comparing the performance of collection selection algorithms," *ACM Transactions on Information Systems*, vol. 21, no. 4, pp. 412–456, ISSN 1046-8188, 2003.

[204] A. Powell, J. French, J. Callan, M. Connell, and C. Viles, "The impact of database selection on distributed searching," in *Belkin et al. [26]*, pp. 232–239. ISBN 1-58113-226-3.

[205] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*. New York, NY: Cambridge University Press, 1988. ISBN 0-521-35465-X.

[206] S. Raghavan and H. García-Molina, "Crawling the hidden web," in *Apers et al. [6]*, pp. 129–138. ISBN 1-55860-804-4.

[207] Y. Rasolofo, F. Abbaci, and J. Savoy, "Approaches to collection selection and results merging for distributed information retrieval," in *Paques et al. [200]*, pp. 191–198. ISBN 1-58113-436-3.

[208] Y. Rasolofo, D. Hawking, and J. Savoy, "Result merging strategies for a current news metasearcher," *Information Processing and Management*, vol. 39, no. 4, pp. 581–609, ISSN 0306-4573, 2003.

[209] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Schenker, "A scalable content-addressable network," in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, pp. 161–172, San Diego, CA: ACM, 2001.

[210] M. Renda and U. Straccia, "Metasearch: rank vs. score based rank list fusion methods (without training data)," Technical report, Instituto di Elaborazione della Informazione — C.N.R., Pisa, Italy, 2002.

[211] M. Renda and U. Straccia, "Web metasearch: Rank vs. score based rank aggregation methods," in *Proceedings of the ACM Symposium on Applied Computing*, (G. Lamont, H. Haddad, G. Papadopoulos, and B. Panda, eds.), pp. 841–846, Melbourne, FL, 2003. ISBN 1-58113-624-2.

[212] S. Robertson, "Relevance weighting of search terms," *Journal of the American Society for Information Sciences*, vol. 27, no. 3, pp. 129–146, 1976.

[213] S. Robertson, "The probability ranking principle in IR," in *Readings in Information Retrieval*, pp. 281–286, Morgan Kaufmann, 1997. ISBN 1-55860-454-5.

[214] S. Robertson and S. Walker, "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (B. Croft and K. Rijsbergen, eds.), pp. 232–241, Dublin, Ireland: ACM/Springer, 1994. ISBN 3-540-19889-X.

[215] G. Salton, E. Fox, and E. Voorhees, "A comparison of two methods for boolean query relevance feedback," Technical report, Cornell University, Ithaca, NY, 1983.

[216] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill, 1986. ISBN 0070544840.

[217] J. Savoy, A. Calvé, and D. Vrajitoru, "Information retrieval systems for large document collections," in *Proceedings of the Fifth Text REtrieval Conference*, (E. Voorhees and D. Harman, eds.), pp. 489–502, Gaithersburg, MD, 1996. NIST Special Publication.

[218] F. Schumacher and R. Eschmeyer, "The estimation of fish populations in lakes and ponds," *Journal of the Tennesse Academy of Science*, vol. 18, pp. 228–249, 1943.

[219] E. Selberg and O. Etzioni, "Multi-service search and comparison using the metacrawler," in *Proceedings of the Fourth International Conference on World Wide Web*, Boston, MA: Oreilly, 1995. ISBN 978-1-56592-169-6.

[220] E. Selberg and O. Etzioni, "The MetaCrawler architecture for resource aggregation on the web," *IEEE Expert*, vol. 12, no. 1, pp. 8–14, ISSN 0885-9000, 1997.

[221] E. Selberg and O. Etzioni, "The MetaCrawler architecture for resource aggregation on the web," *IEEE Expert, January–February*, pp. 11–14, ISSN 0885-9000, 1997.

[222] J. Seo and B. Croft, "Blog site search using resource selection," in *Shanahan et al. [223]*, pp. 1053–1062. ISBN 978-1-59593-991-3.

[223] J. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D. Evans, A. Kolcz, K. Choi, and A. Chowdhury, eds., *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*. Napa Valley, CA, 2008. ISBN 978-1-59593-991-3.

[224] V. Shen, C. Saito, C. Lyu, and M. Zurko, eds., *Proceedings of the 10th International Conference on World Wide Web*. Hong Kong, China: ACM, 2001. ISBN 1-58113-348-0.

[225] Y. Shen and D. Lee, "A meta-search method reinforced by cluster descriptors," in *Proceedings of the Second International Conference on Web Information Systems Engineering*, (M. Özsu, H. Schek, K. Tanaka, Y. Zhang, and Y. Kambayashi, eds.), pp. 125–132, Kyoto, Japan: IEEE Computer Society, 2001. ISBN 0-7695-1393-X.

[226] M. Shokouhi, "Central-rank-based collection selection in uncooperative distributed information retrieval," in *Proceedings of the 29th European Conference on Information Retrieval Research, vol. 4425 of Lecture Notes in Computer Science*, (G. Amati, C. Carpineto, and G. Romano, eds.), pp. 160–172, Rome, Italy: Springer, 2007.

[227] M. Shokouhi, "Segmentation of search engine results for effective data-fusion," in *Proceedings of the 29th European Conference on Information Retrieval Research,* vol. 4425 of *Lecture Notes in Computer Science*, (G. Amati, C. Carpineto, and G. Romano, eds.), pp. 185–197, Rome, Italy: Springer, 2007.

[228] M. Shokouhi, M. Baillie, and L. Azzopardi, "Updating collection representations for federated search," in *Kraaij et al. [151]*, pp. 511–518. ISBN 978-1-59593-597-7.

[229] M. Shokouhi, F. Scholer, and J. Zobel, "Sample sizes for query probing in uncooperative distributed information retrieval," in *Proceedings of Eighth Asia Pacific Web Conference*, (X. Zhou, J. Li, H. Shen, M. Kitsuregawa, and Y. Zhang, eds.), pp. 63–75, Harbin, China, 2006. ISBN 3-540-31142-4.

[230] M. Shokouhi, P. Thomas, and L. Azzopardi, "Effective query expansion for federated search," in *Allan et al. [3]*, pp. 427–434. ISBN 978-1-60558-483-6.

[231] M. Shokouhi and J. Zobel, "Federated text retrieval from uncooperative overlapped collections," in *Kraaij et al. [151]*, pp. 495–502. ISBN 978-1-59593-597-7.

[232] M. Shokouhi and J. Zobel, "Robust result merging using sample-based score estimates," *ACM Transactions on Information Systems*, vol. 27, no. 3, pp. 1–29, ISSN 1046-8188, 2009.

[233] M. Shokouhi, J. Zobel, and Y. Bernstein, "Distributed text retrieval from overlapping collections," in *Proceedings of the 18th Australasian Database Conference,* vol. 63 of *CRPIT*, (J. Bailey and A. Fekete, eds.), pp. 141–150, Ballarat, Australia: ACS, 2007.

[234] M. Shokouhi, J. Zobel, and Y. Bernstein, "Distributed text retrieval from overlapping collections," in *Proceedings of the Australasian Database Conference*, (J. Bailey and A. Fekete, eds.), pp. 141–150, Ballarat, Australia, 2007.

[235] M. Shokouhi, J. Zobel, F. Scholer, and S. Tahaghoghi, "Capturing collection size for distributed non-cooperative retrieval," in *Efthimiadis et al. [87]*, pp. 316–323. ISBN 1-59593-369-7.

[236] M. Shokouhi, J. Zobel, S. Tahaghoghi, and F. Scholer, "Using query logs to establish vocabularies in distributed information retrieval," *Information Processing and Management*, vol. 43, no. 1, pp. 169–180, ISSN 0306-4573, 2007.

[237] X. M. Shou and M. Sanderson, "Experiments on data fusion using headline information," in *Järvelin et al. [142]*, pp. 413–414. ISBN 1-58113-561-0.

[238] S. Shushmita, H. Joho, M. Lalmas, and R. Villa, eds., *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management.* Toronto, Canada, 2010.

[239] L. Si, "Federated search of text search engines in uncooperative environments," PhD thesis, Carnegie Mellon University, 2006.

[240] L. Si and J. Callan, "The effect of database size distribution on resource selection algorithms," in *Callan et al. [44]*, pp. 31–42. ISBN 3-540-20875-5.

[241] L. Si and J. Callan, "Modeling search engine effectiveness for federated search," in *Marchionini et al. [181]*, pp. 83–90. ISBN 1-59593-034-5.

[242] L. Si and J. Callan, "Relevant document distribution estimation method for resource selection," in *Clarke et al. [56]*, pp. 298–305. ISBN 1-58113-646-3.

[243] L. Si and J. Callan, "Using sampled data and regression to merge search engine results," in *Järvelin et al. [142]*, pp. 19–26. ISBN 1-58113-561-0.

[244] L. Si and J. Callan, "A semisupervised learning method to merge search engine results," *ACM Transactions on Information Systems*, vol. 21, no. 4, pp. 457–491, ISSN 1046-8188, 2003.

[245] L. Si and J. Callan, "Unified utility maximization framework for resource selection," in *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*, (D. Grossman, L. Gravano, C. Zhai, O. Herzog, and D. Evans, eds.), pp. 32–41, Washington, DC, 2004. ISBN 1-58113-874-1.

[246] L. Si and J. Callan, "CLEF2005: multilingual retrieval by combining multiple multilingual ranked lists," in *The Sixth Workshop of the Cross-Language Evaluation Forum*, Vienna, Austria, 2005. URL %%%%`http://www.cs.purdue.edu/homes/lsi/publications.htm`.

[247] L. Si, J. Callan, S. Cetintas, and H. Yuan, "An effective and efficient results merging strategy for multilingual information retrieval in federated search environments," *Information Retrieval*, vol. 11, no. 1, pp. 1–24, 2008.

[248] L. Si, R. Jin, J. Callan, and P. Ogilvie, "A language modeling framework for resource selection and results merging," in *Kalpakis et al. [144]*, pp. 391–397. ISBN 1-58113-492-4.

[249] A. Smeaton and F. Crimmins, "Using a data fusion agent for searching the WWW," in *Selected papers from the Sixth International Conference on World Wide Web*, (P. Enslow, M. Genesereth, and A. Patterson, eds.), Santa Clara, CA: Elsevier, 1997. Poster Session.

[250] M. Sogrine, T. Kechadi, and N. Kushmerick, "Latent semantic indexing for text database selection," in *Proceedings of the SIGIR 2005 Workshop on Heterogeneous and Distributed Information Retrieval*, pp. 12–19, 2005. URL %%%%`http://hdir2005.isti.cnr.it/index.html`.

[251] A. Spink, B. Jansen, C. Blakely, and S. Koshman, "A study of results overlap and uniqueness among major web search engines," *Information Processing and Management*, vol. 42, no. 5, pp. 1379–1391, ISSN 0306-4573, 2006.

[252] I. Stoica, R. Morris, D. Liben-Nowell, D. Karger, M. Kaashoek, F. Dabek, and H. Balakrishnan, "Chord: A scalable peer-to-peer lookup protocol for internet applications," *IEEE/ACM Transactions on Networking*, vol. 11, no. 1, pp. 17–32, ISSN 1063-6692, 2003.

[253] A. Sugiura and O. Etzioni, "Query routing for web search engines: Architectures and experiments," in *Herman and Vezza [132]*, pp. 417–429. ISBN 1-930792-01-8.

[254] P. Thomas, "Generalising multiple capture-recapture to non-uniform sample sizes," in *Myaeng et al. [192]*, pp. 839–840. ISBN 978-1-60558-164-4.

[255] P. Thomas, "Server characterisation and selection for personal metasearch," PhD thesis, Australian National University, 2008.

[256] P. Thomas and D. Hawking, "Evaluating sampling methods for uncooperative collections," in *Kraaij et al. [151]*, pp. 503–510. ISBN 978-1-59593-597-7.

[257] P. Thomas and D. Hawking, "Experiences evaluating personal metasearch," in *Proceedings of the second international Symposium on Information Interaction in Context*, pp. 136–138, London, UK: ACM, 2008. ISBN 978-1-60558-310-5.

[258] P. Thomas and D. Hawking, "Server selection methods in personal metasearch: a comparative empirical study," *Information Retrieval*, vol. 12, no. 5, pp. 581–604, ISSN 1386-4564, 2009.

[259] P. Thomas and M. Shokouhi, "SUSHI: scoring scaled samples for server selection," in *Allan et al. [3]*, pp. 419–426. ISBN 978-1-60558-483-6.

[260] G. Towell, E. Voorhees, K. Narendra, and B. Johnson-Laird, "Learning collection fusion strategies for information retrieval," in *Proceedings of The 12th International Conference on Machine Learning*, (A. Prieditis and S. Russell, eds.), pp. 540–548, Lake Tahoe, CA, 1995. ISBN 1-55860-377-8.

[261] T. Tsikrika and M. Lalmas, "Merging techniques for performing data fusion on the web," in *Paques et al. [200]*, pp. 127–134. ISBN 1-58113-436-3.

[262] H. Turtle, "Inference networks for document retrieval," PhD thesis, University of Massachusetts, 1991.

[263] H. Turtle and B. Croft, "Inference networks for document retrieval," in *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (J. Vidick, ed.), pp. 1–24, Brussels, Belgium, 1990. ISBN 0-89791-408-2.

[264] C. Vogt, "Adaptive combination of evidence for information retrieval," PhD thesis, University of California, San Diego, 1999.

[265] C. Vogt and G. Cottrell, "Fusion via a linear combination of scores," *Information Retrieval*, vol. 1, no. 3, pp. 151–173, ISSN 1386-4564, 1999.

[266] E. Voorhees, N. Gupta, and B. Johnson-Laird, "Learning collection fusion strategies," in *Fox et al. [91]*, pp. 172–179. ISBN 0-89791-714-6.

[267] E. Voorhees and R. Tong, "Multiple search engines in database merging," in *Proceedings of the Second ACM Conference on Digital Libraries*, (R. Allen and E. Rasmussen, eds.), pp. 93–102, Philadelphia, PA, 1997. ISBN 0-89791-868-1.

[268] Y. Wang and D. DeWitt, "Computing PageRank in a distributed internet search engine system," in *Proceedings of the 30th International Conference on Very Large Data Bases*, (M. Nascimento, M. Özsu, D. Kossmann, R. Miller, J. Blakeley, and K. Schiefer, eds.), pp. 420–431, Toronto, Canada: Morgan Kaufmann, 2004. ISBN 0-12-088469-0.

[269] C. Williamson, M. Zurko, P. Patel-Schneider, and P. Shenoy, eds., *Proceedings of the 16th International Conference on World Wide Web*. Alberta, Canada: ACM, 2007. ISBN 978-1-59593-654-7.

[270] S. Wu and F. Crestani, "Multi-objective resource selection in distributed information retrieval," in *Kalpakis et al. [144]*, pp. 1171–1178. ISBN 1-58113-492-4.

[271] S. Wu and F. Crestani, "Distributed information retrieval: A multi-objective resource selection approach," *International Journal of Uncertainty, Fuzziness Knowledge-Based Systems*, vol. 11, no. supp01, pp. 83–99, ISSN 0218-4885, 2003.

[272] S. Wu and F. Crestani, "Shadow document methods of resutls merging," in *Proceedings of the ACM symposium on Applied computing*, (H. Haddad, A. Omicini, R. Wainwright, and L. Liebrock, eds.), pp. 1067–1072, Nicosia, Cyprus, 2004. ISBN 1-58113-812-1.

[273] S. Wu and S. McClean, "Performance prediction of data fusion for information retrieval," *Information Processing and Management*, vol. 42, no. 4, pp. 899–915, ISSN 0306-4573, 2006.

[274] Z. Wu, W. Meng, C. Yu, and Z. Li, "Towards a highly-scalable and effective metasearch engine," in *Shen et al. [224]*, pp. 386–395. ISBN 1-58113-348-0.

[275] J. Xu and J. Callan, "Effective retrieval with distributed collections," in *Croft et al. [73]*, pp. 112–120. ISBN 1-58113-015-5.

[276] J. Xu and B. Croft, "Cluster-based language models for distributed retrieval," in *Gey et al. [106]*, pp. 254–261. ISBN 1-58113-096-1.

[277] J. Xu and B. Croft, "Query expansion using local and global document analysis," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (H. Frei, D. Harman, P. Schäuble, and R. Wilkinson, eds.), pp. 4–11, Zurich, Switzerland, 1996. ISBN 0-89791-792-8.

[278] J. Xu, S. Wu, and X. Li, "Estimating collection size with logistic regression," in *Kraaij et al. [151]*, pp. 789–790. ISBN 978-1-59593-597-7.

[279] H. Yang and M. Zhang, "Ontology-based resource descriptions for distributed information sources," in *Proceedings of the Third International Conference on Information Technology and Applications,* vol. I, (X. He, T. Hintza, M. Piccardi, Q. Wu, M. Huang, and D. Tien, eds.), pp. 143–148, Sydney, Australia: IEEE Computer Society, 2005. ISBN 0-7695-2316-1.

[280] H. Yang and M. Zhang, "Two-stage statistical language models for text database selection," *Information Retrieval*, vol. 9, no. 1, pp. 5–31, ISSN 1386-4564, 2006.

[281] C. Yu, K. Liu, W. Meng, Z. Wu, and N. Rishe, "A methodology to retrieve text documents from multiple databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 6, pp. 1347–1361, ISSN 1041-4347, 2002.

[282] C. Yu, W. Meng, K. Liu, W. Wu, and N. Rishe, "Efficient and effective metasearch for a large number of text databases," in *Gauch [102]*, pp. 217–224. ISBN 1-58113-1461.

[283] C. Yu, W. Meng, W. Wu, and K. Liu, "Efficient and effective metasearch for text databases incorporating linkages among documents," *SIGMOD Records*, vol. 30, no. 2, pp. 187–198, ISSN 0163-5808, 2001.

[284] B. Yuwono and D. Lee, "WISE: A world wide web resource database system," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 4, pp. 548–554, ISSN 1041-4347, 1996.

[285] B. Yuwono and D. Lee, "Server ranking for distributed text retrieval systems on the internet," in *Proceedings of the Fifth International Conference on Database Systems for Advanced Applications,* vol. 6 of *Advanced Database Research and Development Series*, (R. Topor and K. Tanaka, eds.), pp. 41–50, Melbourne, Australia: World Scientific, 1997. ISBN 981-02-3107-5.

[286] O. Zamir and O. Etzioni, "Grouper: a dynamic clustering interface to web search results," *Computer Networks and ISDN Systems*, vol. 31, no. 11–16, pp. 1361–1374, ISSN 1389-1286, 1999.

[287] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, "Fully automatic wrapper generation for search engines," in *Ellis and Hagino [88]*, pp. 66–75. ISBN 1-59593-046-9.

[288] H. Zhao, W. Meng, and C. Yu, "Automatic extraction of dynamic record sections from search engine result pages," in *Proceedings of the 30th International Conference on Very Large Data Bases*, (U. Dayal, K.-Y. Whang, D. B. Lomet, G. Alonso, G. M. Lohman, M. L. Kersten, S. K. Cha, and Y.-K. Kim, eds.), pp. 989–1000, Seoul, Korea: Morgan Kaufmann, 2006. ISBN 1-59593-385-9.

[289] H. Zhao, W. Meng, and C. Yu, "Mining templates from search result records of search engines," in *Proceedings of the 13th Annual International ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, (P. Berkhin, R. Caruana, and X. Wu, eds.), pp. 884–893, 2007: San Jose, California, USA. ISBN 978-1-59593-609-7.

[290] J. Zobel, "Collection selection via lexicon inspection," in *Proceedings of the Australian Document Computing Symposium*, (P. Bruza, ed.), pp. 74–80, Melbourne, Australia, 1997.