Distributed Information Retrieval and Applications

Fabio Crestani and Ilya Markov

University of Lugano, Via G. Buffi 13, 6900, Lugano, Switzerland
{fabio.crestani,ilya.markov}@usi.ch

Abstract. Distributed Information Retrieval (DIR) is a generic area of research that brings together techniques, such as resource selection and results aggregation, dealing with data that, for organizational or technical reasons, cannot be managed centrally. Existing and potential applications of DIR methods vary from blog retrieval to aggregated search and from multimedia and multilingual retrieval to distributed Web search. In this tutorial we briefly discuss main DIR phases, that are resource description, resource selection, results merging and results presentation. The main focus is made on applications of DIR techniques: blog, expert and desktop search, aggregated search and personal meta-search, multimedia and multilingual retrieval. We also discuss a number of potential applications of DIR techniques, such as distributed Web search, enterprise search and aggregated mobile search.

1 Introduction

Distributed Information Retrieval (DIR), also known as Federated Search or Federated IR, concerns with aggregating multiple searchable sources of information under a single interface [6,23]. DIR consists of the following phases: (i) resource description or representation, where a high-level description is built for each federated source. (ii) Server/resource selection, where, given a user's query, several relevant sources are selected for further processing. (iii) Results merging or aggregation, where the results obtained from selected sources are combined into a single result list. (iv) Results presentation, where the obtained results are grouped and positioned on a result page.

However, modern applications of the standard DIR techniques usually have a different set of assumptions and limitations [29]. For instance, aggregated search works in highly cooperative environments and does not need to merge results into a single list, blog and expert search do not require description and merging phases, while multilingual retrieval requires additional steps of query and document translation. In this tutorial we discuss the standard DIR techniques and show how they can be adapted and applied to various IR problems, such as blog distillation and desktop search, aggregated search and distributed Web search, multimedia and multilingual retrieval.

2 Distributed Information Retrieval

In this section we briefly discuss main DIR phases, such as resource description, resource selection, results merging and results presentation.

Resource Description. In the offline phase a high-level description is built for each federated source. The description may include a full content of a source (or only a sample of its documents in uncooperative environments [7]), term and document statistics,

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 865–868, 2013.

metadata (if available) and other descriptors of the source's content. The descriptions of all federated sources are managed centrally by a DIR broker and are used for subsequent phases, such as resource selection and results merging.

Resource Selection. Given a user's query and the sources' descriptions, the DIR broker selects the most relevant sources for the query. First generation resource selection techniques, also known as large document approaches, represent each source as a concatenation of its documents. The obtained large documents are ranked using standard IR techniques, such as adapted INQUERY in CORI [5] and language modeling in [33]. Second generation or small document approaches use a centralized sample index of documents and rank sources based on the number and the position of their documents in a centralized ranking (eg. ReDDE [26], CRCS [22] and others [15,18,20,32]). Finally, classification-based resource selection combines the above approaches and a number of other query- and corpus-based features in a machine learning framework [1,12].

Results Merging and Score Normalization. The user's query is forwarded to the selected sources and the retrieved source-specific results are merged into a single list using results merging and score normalization methods. Results merging techniques use sources' descriptions either implicitly through resource selection, like CORI [5,17], or explicitly like SSL [25] and SAFE [24]. Score normalization methods do not use descriptions, but require document relevance scores to be provided by federated sources [16].

Results Presentation. Instead of merging results into a single list or in addition to that, the results may be presented to a user in various ways, eg. blended, tabbed, side-by-side, etc. [28,31].

3 Applications

In this section we discuss the applications and adaptations of DIR techniques to various IR tasks.

Applications of Resource Selection. Resource selection is probably the most widely used DIR technique. For example, in blog distillation the most relevant blogs need to be retrieved for a user's query. Since each blog is a collection of posts, blog distillation can be treated as a resource selection problem and both small and large document approaches can be applied to solve it [11,21]. The same idea is applicable to expert search if each expert is considered as a collection of documents that he/she produced. In desktop search each document type may be treated as a separate documents source and resource selection may be performed to select the most relevant one [13].

Aggregated Search. Aggregated search augments Web search results with the results of several vertical searches [3]. It can be seen as a variation of DIR, where aggregated verticals are managed by the same content provider and, therefore, are fully cooperative. Moreover, verticals are highly heterogeneous in content and types of media. Still most of DIR steps are present here. First, resource description may be performed for efficiency reasons [3]. Second, vertical selection can be seen as a variation of resource selection that accounts for multiple media types and where no vertical can be selected [3]. Finally, the vertical-specific results need to be blended into the Web results [2].

Other Applications. Multilingual retrieval combines documents written in multiple languages and, therefore, benefits from results merging methods [27]. Multimedia distributed

digital libraries can use a standard DIR architecture [8]. A personal meta-search system provides a user with a possibility of searching the Web and personal content with a single interface using DIR techniques [30]. Similarly, federated Web search provides a low-cost solution for aggregating multiple searchable Web sites within a single interface [19].

4 Potential Applications

In this section we discuss distributed Web search, enterprise search and aggregated search on mobile devices as potential applications of DIR methods.

Distributed Web Search. Site selection in distributed Web search is a specific resource selection problem, where search sites are not autonomous, resources and algorithms are homogeneous and the distribution of content is managed explicitly [9]. Currently site selection is performed based on a predicted query performance [4] or a thresholding algorithm [10]. Although some work has been done on the feasibility of DIR methods for topically partitioned collections [14], the applicability of existing resource selection techniques to distributed Web search and the development of novel specific methods still needs to be addressed.

Enterprise Search. Enterprise search involves different types of document sources used within a company as well as a number of external searchable collections that need to be aggregated into a single company-wide search system. DIR techniques such as resource selection and results presentation may help in solving this problem [29].

Aggregated Mobile Search. Aggregated search on mobile devices has a number of unique features, such as a rich context, a specific interaction of a user with search results and technical limitations of mobile devices. The applicability of existing DIR techniques and the development of novel task-specific methods for aggregated mobile search are yet to be addressed.

References

1. Arguello, J., Callan, J., Diaz, F.: Classification-based resource selection. In: Proceedings of CIKM, pp. 1277–1286. ACM (2009)
2. Arguello, J., Diaz, F., Callan, J.: Learning to aggregate vertical results into web search results. In: Proceedings of CIKM, pp. 201–210 (2011)
3. Arguello, J., Diaz, F., Callan, J., Crespo, J.F.: Sources of evidence for vertical selection. In: Proceedings of SIGIR, pp. 315–322 (2009)
4. Baeza-Yates, R., Murdock, V., Hauff, C.: Efficiency trade-offs in two-tier web search systems. In: Proceedings of SIGIR, pp. 163–170 (2009)
5. Callan, J.P., Lu, Z., Croft, W.B.: Searching distributed collections with inference networks. In: Proceedings of SIGIR, pp. 21–28 (1995)
6. Callan, J.: Advances in Information Retrieval. In: Distributed Information Retrieval, vol. ch. 5, pp. 127–150. Kluwer Academic Publishers (2000)
7. Callan, J., Connell, M.: Query-based sampling of text databases. ACM Transactions of Information Systems 19(2), 97–130 (2001)
8. Callan, J., Crestani, F., Nottelmann, H., Pala, P., Shou, X.M.: Resource selection and data fusion in multimedia distributed digital libraries. In: Proceedings of SIGIR, pp. 363–364 (2003)
9. Cambazoglu, B.B., Plachouras, V., Baeza-Yates, R.: Quantifying performance and quality gains in distributed web search engines. In: Proceedings of SIGIR, pp. 411–418 (2009)
10. Cambazoglu, B.B., Varol, E., Kayaaslan, E., Aykanat, C., Baeza-Yates, R.: Query forwarding in geographically distributed search engines. In: Proceedings of SIGIR, pp. 90–97 (2010)

11. Elsas, J.L., Arguello, J., Callan, J., Carbonell, J.G.: Retrieval and feedback models for blog feed search. In: Proceedings of SIGIR, pp. 347–354 (2008)
12. Hong, D., Si, L., Bracke, P., Witt, M., Juchcinski, T.: A joint probabilistic classification model for resource selection. In: Proceedings of SIGIR, pp. 98–105 (2010)
13. Kim, J., Croft, W.B.: Ranking using multiple document types in desktop search. In: Proceedings of SIGIR. pp. 50–57 (2010)
14. Kulkarni, A., Callan, J.: Document allocation policies for selective searching of distributed indexes. In: Proceedings of CIKM, pp. 449–458 (2010)
15. Markov, I.: Modeling document scores for distributed information retrieval. In: Proceedings of SIGIR, pp. 1321–1322 (2011)
16. Markov, I., Arampatzis, A., Crestani, F.: Unsupervised linear score normalization revisited. In: Proceedings of SIGIR, pp. 1161–1162 (2012)
17. Markov, I., Arampatzis, A., Crestani, F.: On CORI results merging. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Agichtein, S.R.E., Segalovich, I., Yilmaz, E. (eds.) ECIR 2013. LNCS, vol. 7814, pp. 736–739. Springer, Heidelberg (2013)
18. Markov, I., Azzopardi, L., Crestani, F.: Reducing the uncertainty in resource selection. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Agichtein, S.R.E., Segalovich, I., Yilmaz, E. (eds.) ECIR 2013. LNCS, vol. 7814, pp. 500–511. Springer, Heidelberg (2013)
19. Nguyen, D., Demeester, T., Trieschnigg, D., Hiemstra, D.: Federated search in the wild: the combined power of over a hundred search engines. In: Proceedings of CIKM, pp. 1874–1878 (2012)
20. Paltoglou, G., Salampasis, M., Satratzemi, M.: Integral based source selection for uncooperative distributed information retrieval environments. In: Proceedings of the ACM LSDS-IR Workshop, pp. 67–74 (2008)
21. Seo, J., Croft, W.B.: Blog site search using resource selection. In: Proceedings of CIKM, pp. 1053–1062 (2008)
22. Shokouhi, M.: Central-Rank-Based Collection Selection in Uncooperative Distributed Information Retrieval. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECIR 2007. LNCS, vol. 4425, pp. 160–172. Springer, Heidelberg (2007)
23. Shokouhi, M., Si, L.: Federated search. Foundations and Trends in Information Retrieval 5, 1–102 (2011)
24. Shokouhi, M., Zobel, J.: Robust result merging using sample-based score estimates. ACM Transactions of Information Systems 27(3), 1–29 (2009)
25. Si, L., Callan, J.: Using sampled data and regression to merge search engine results. In: Proceedings of SIGIR, pp. 19–26 (2002)
26. Si, L., Callan, J.: Relevant document distribution estimation method for resource selection. In: Proceedings of SIGIR, pp. 298–305 (2003)
27. Si, L., Callan, J., Cetintas, S., Yuan, H.: An effective and efficient results merging strategy for multilingual information retrieval in federated search environments. Information Retrieval 11(1), 1–24 (2008)
28. Sushmita, S., Joho, H., Lalmas, M., Villa, R.: Factors affecting click-through behavior in aggregated search interfaces. In: Proceedings of CIKM, pp. 519–528 (2010)
29. Thomas, P.: To what problem is distributed information retrieval the solution? Journal of the American Society for Information Science and Technology 63(7), 1471–1476 (2012)
30. Thomas, P., Hawking, D.: Server selection methods in personal metasearch: a comparative empirical study. Information Retrieval 12(5), 581–604 (2009)
31. Thomas, P., Noack, K., Paris, C.: Evaluating interfaces for government metasearch. In: Proceedings of IIiX, pp. 65–74 (2010)
32. Thomas, P., Shokouhi, M.: Sushi: scoring scaled samples for server selection. In: Proceedings of SIGIR, pp. 419–426 (2009)
33. Xu, J., Croft, W.B.: Cluster-based language models for distributed retrieval. In: Proceedings of SIGIR, pp. 254–261 (1999)