Distributed Information Retrieval using Keyword Auctions

Djoerd Hiemstra
University of Twente
http://www.cs.utwente.nl/~hiemstra

Abstract This report motivates the need for large-scale distributed approaches to information retrieval, and proposes solutions based on keyword auctions.

1 Introduction

After the burst of the dot-com bubble in the autumn of 2001, the World Wide Web has gone through some remarkable changes in its organizational structure. Consumers of data and content are increasingly taking the role of *producers* of data and content, thereby threatening traditional publishers. A well known example is the Wikipedia encyclopedia, which is written entirely by its (non-professional) users on a voluntary basis, while still rivaling a traditional publisher like Britannica on-line in both size and quality [31]. Similarly, in SourceForge, communities of open source software developers collaboratively create new software thereby rivaling software vendors like Microsoft; Blogging turned the internet consumers of news into news providers; Kazaa and related peer-to-peer platforms like BitTorrent and E-mule turned anyone who downloads a file automatically into contributors of files; Flickr turned users into contributors of visual content, but also into indexers of that content by social tagging, etc. Communities of users operate by trusting each other as co-developers and contributors, without the need for strict rules. There is however one major internet application for which communities only play a minor role. One of the web's most important applications – if not *the* most important application – is *search*. Internet search is almost exclusively run by a small number of powerful companies that dominate the search market: Google, Yahoo, and Microsoft. In contrast to traditional centralized search, where a centralized body like Google or Yahoo is in full control, a community-run search engine would consist of many small search engines that collaboratively provide the search service. The overall aim of this proposal is to develop models and approaches that enable such a community-run search service, i.e.:

> *To distribute internet search functionality in such a way that communities of users and/or federations of small search systems provide search services in a collaborative way.*

This report motivates the need for distributed information retrieval and briefly addresses related work in Section 2. Section 3 proposes a novel solution to the problem using the analogy of keyword auctions. Section 4 concludes this paper.

2 The need for large-scale distributed search

Companies such as Google and Yahoo control every aspect of internet search: 1) collecting content (i.e. crawling), 2) indexing the content, and of course 3) searching the content. Typically in centralized search, a central monolithic database is filled by copying, usually by crawling, all data that needs to be indexed and searched. Note that the centralized search service might run on a large cluster of work stations, and also note that the index might be replicated in several distinct data centers: The point is that the *control* of data is centralized. Running a centralized web search service is getting increasingly more difficult because of the growth of the number of web pages. Already in 1999, it was estimated that no search engine indexes more than 16% of the visible web [48]. Worse still, a large part of the web, the invisible web or *deep web* which size is estimated to be about 500 times bigger than the visible web [12], cannot be indexed by web crawlers at all (for instance dynamic pages that are returned in response to filling in a web form). A third major problem is the *freshness* of the index. Changes in web pages and other web data can only be observed if they are visited by the web crawler. A fourth problem is the increasing availability of *structured data* and the need to search this data [54]. Structured data is provided via web forms and as such not available for web crawlers. Structured data is also provided by many emerging XML standards such as RSS or MPEG-7. Maybe the biggest problem of centralized search is that of *monopoly power* of the major search engines like Google. Whoever controls search, controls how we see the world, what information we read, what products we buy, etc. For this reason, the French president Jacques Chirac argued that the EU should fund a European internet search engine that rivals Yahoo and Google by providing competition to what he sees as "the threat of Anglo-Saxon cultural imperialism" [83]. Whereas the support of a search engine like Yahoo for French seems to be adequate (http://fr.yahoo.com), indeed, there is ample support for commercially less interesting languages, such as Polish or Persian.

2.1 Scientific challenges and related work

In *Distributed Information Retrieval* [18], a centralized broker provides access to multiple text databases by sending the user's query to the databases that are most likely to satisfy the user's need for information. The broker collects the results from each database and merges those in one final result. To achieve this, the distributed information retrieval system has to address the following: 1) *resource description*, i.e., what are the contents of each text database? 2) *resource selection*, i.e., given a query, what text databases should be contacted?, and 3) *results merging*, i.e., how to integrate the results returned by the text databases into one coherent ranked list?

*Resource description*s are usually not obtained directly from the local databases because the databases are either uncooperative or, if they are cooperative, their descriptions might not be trusted. Instead, resource descriptions can be obtained by so-called query-based sampling [19], [67], which involves sending a number of queries to each local database to obtain a sample of the documents that are indexed by that database. Queries are chosen from previously retrieved documents, from query logs [23], [68], or from some topic hierarchy [32]. Sampling is terminated when the acquired sample is large enough to be useful in the resource selection phase, often after several hundreds of sampled documents [4], [7]. Challenges are to reduce network load (downloading a few hundred documents for every server might constitute a substantial part of the full collection), and to keep the resource descriptions up-to-date when the contents of the local databases change [67].

Resource selection uses techniques similar to ordinary document search. A first family of methods concatenates the sampled documents of each local database together to form a compound document [69], [79]. This way server selection can be done by standard information retrieval methods, i.e., by ranking the compound documents given a query. A second family of methods ranks the separate documents in the complete sampled set to predict which local database needs to be selected [69] [73]. A third family of methods is based on distributing the information of the broker over the local systems as well. In these peer-to-peer systems, every local system is both broker and local database, an architecture that fits well with community-based applications [64]. Peer-to-peer architectures for information retrieval tasks have been well-studied, for instance PlanetP [27], pSearch [77], Odissea [71], Minerva [11], and Alvis [53]. Such implementations work in theory, but cause high bandwidth consumption in practice because in large networks, many of the important terms occur in almost every peer [49], [82]. The main challenge of resource selection is keeping the amount of data to be sent (i.e. the number of databases to be contacted) low, while still obtaining good quality search results [57], for instance by modeling both quality and efficiency [60]. Another challenge is to find non-content criteria for resource selection, like the average quality of a page from the database or the estimated size of the database [73] [79].

Results merging often involves normalization of the relevance scores of the retrieved documents [20], [69]. If scores are not available, or if the score from the local databases cannot be trusted, which is often the case in real life search applications, merging might be based on the ranks of the retrieved documents [4], [62]. In this case, the sampled documents mentioned above can be used to (re-)calculate the score locally, but this requires a significant sample of each local database. One of the main challenges of result merging is finding objective criteria for merging that do not need document scores or samples of downloaded documents.

Distributed information retrieval is well-studied, and it has many potential benefits over a centralized approach. However, research did not result in practical solutions for large scale search problems, and the challenges mentioned above have not been solved sufficiently for distributed information retrieval to be used in real life. In fact, the research topic is still very much alive [6]. We believe there is an underlying reason for the failure of distributed architectures until now. This underlying challenge was nicely formulated by Sergey Brin and Larry Page when they introduced Google [16]:

> *Of course a distributed system (…) will often be the most efficient and elegant technical solution for indexing, but it seems difficult to convince the world to use these systems because of the high administration costs of setting up large numbers of installations. Of course, it is quite likely that reducing the administration cost drastically is possible. If that happens, and everyone starts running a distributed indexing system, searching would certainly improve drastically.*

So, the main challenge is to drastically reduce the administration costs of distributed information retrieval. Keyword auctions have the potential to meet this challenge.

2.2 Keyword Auctions

Search companies like Google and Yahoo did not only revolutionize search, they also changed the way businesses advertise on the web [56]. The on-line advertisement system used by Google, called AdWords, operates as an auction of keywords as follows. Advertisers bid on individual keywords, while indicating a limit for a daily budget. Advertisers pay per click, that is, they pay only for a displayed advertisement if a user actually clicks it. Now, Google will display for each query the advertisements from which it expects the greatest revenue. Note that this is not necessarily the advertisement of the highest bidder, because it might have a click-through rate that is twice as low as the next highest bidder, who bids for instance 80% of the highest bidder. (In practice, they pay the amount of the highest bidder just below them, i.e., the maximum bid of all bids that were lower than the advertisers bid.) Keyword auctions like AdWords have been very successful because they allow advertisers to provide very targeted advertisements to customers, and because they are able server a huge number of small-budget advertisers, the so-called “long tail” [61], i.e. small advertisers make up the bulk of world-wide budgets available for advertising. Very targeted resource selection, and serving of a huge number of small local databases is exactly what is needed for community-based distributed information retrieval.

3 Distributed Search using Keyword Auctions

The characteristics of keyword auctions fit especially well with the needs of community-based distributed information retrieval: They provide very targeted resource selection, and they are able to serve a huge number of small local databases. The keyword auctions analogy will be used for distributed information retrieval as follows: Local search providers bid on individual keywords. In turn they pay by serving queries. As in AdWords they have a limited daily budget, i.e., they are able to serve a limited number of queries per day. Similar to the advertisements of AdWords, the broker displays one (or more) results from the local search providers result pages. Like advertisements, these results consist of a page title, a short text snippet, and a URL. As in AdWords, the broker does not blindly present the results from the highest bidder, but it presents those results that optimize the distributed system’s overall quality, measured in for instance the efficiency of the search, the diversity of the answers, the click-through rates, the similarity of the query to the result title, snippet and URL, etc. As in AdWords, local search providers can change their bids whenever they like, for instance only bid on the keyword “Christmas tree” in December, which leads to a very dynamic environment in which changes to local collections are instantly seen by the search system without the need of crawling the data.

3.1 Objectives

The keyword auction analogy calls for a completely new administrative policy of distributed information retrieval. First of all, the process is initialized by the local search providers, so we need an information push scenario (the local search system contacts the broker) next to the information pull scenario that is usually assumed (the broker contacts the local searcher). Unlike AdWords, the scenario will have to include the possibility to bid on any keyword, i.e., some systems might have indexed a significant part of the web and therefore willing to process any query. This is needed to get good coverage of all possible keywords/queries. Second, communication is done only via bidding and the search engine result page texts (i.e. the “texts of the advertisements”), so we will refrain from sampling documents from the local search providers. Third, bidding is done on complex keywords, i.e., keywords that possibly consist of more than one term, so we need text representations with a higher complexity than the usual bag of words [5], or more specifically, a higher complexity than unigram language model representations [34]. The project’s main objectives are:

Objective 1, Automatic bidding strategies An important part of the proposed research will be done into automatic bidding strategies and automatic keyword extraction strategies. Bidding has to be automated to reduce the administrative costs. We have recently shown that the Google AdWords bidding process can be automated for a large part [81], allowing companies to manage AdWords campaigns without the need to monitor their budgets or their competitors on a regular basis. The project will develop and evaluate automatic bidding strategies for distributed information retrieval.

Objective 2, Optimization strategies of the broker Strategies that search engines follow to optimize the advertisements revenue have been studied extensively in the past years.

Lahaie et al. [46] present a detailed description of these results, of which many are adapted from existing game theoretic analyses. Mehta et al. [56] describe the AdWords problem as a generalization of the on-line bipartite matching problem. Edelman et al. [28] and Varian [75] study the generalized second price auction, the charging scheme where the highest bidder pays the amount of the second highest bidder. The project will develop and evaluate keyword auction optimization algorithms for distributed information retrieval that take into account the amount of queries a search provider can process, click-through data, search engine result pages, etc.

Objective 3, Models for Complex keywords A lot of work on modeling in information retrieval is based on the assumption of independence between terms (words) given a document. For instance, most work on language models [26], [35] is based on simple unigram models. Keywords auctions imply the use of complex keywords also used in studies of click-through data [25]. Interestingly, Podnar et al. [63] suggests complex keywords for peer-to-peer search to improve the efficiency resource selection. We recently applied Podnar's *discriminative* or *rare keys* for distributed query processing [74]. The study showed that highly discriminative keys scale well with collection growth, but it also showed worse performance on web data than the previous studies which used data from Reuters. This suggest additional measures have to be taken to make discriminative keys a success, for instance query adaptive indexing [8] and better modeling of the statistics of complex keywords. The project will develop and evaluate models for complex keywords based language models [35] [36], [37], [40], [41].

Objective 4, Search engine result pages instead of document sampling One of the project's goals is to develop an architecture that does not need the downloading of complete documents. Instead of using document scores and statistics from sampled documents, the project intends to use the information that *is* typically contained in the returned results of local search providers, such as a title, a URL and a text snippet. Searches that only use the titles of pages produce high precision results in centralized web search for navigational queries [43]. The proposed project will investigate the use of information from search engine result pages for resource description and selection. Additionally, the project will investigate the use of the diversity of results, and non-content features such as the estimated size of local databases or a local database' static ranking similar to Pagerank [16].

Objective 5, Seamless integration of structured data and annotation schemes with text search Structured data is increasingly available on the web, and according to Madhavan et al. [54] *"the prime example of such data is the deep web"*, i.e., most data is only accessible via web forms and cannot be crawled by centralized search engines. It will however be accessible in a distributed information retrieval scenario. Other examples are Google Base and structure from annotation schemes, for instance the manual annotation of photos in Flickr. In a distributed information retrieval scenario, structure and annotations might be added to the search results by local search providers if they are able to give text search access to their structured data. This would create an instance of a data*space* system [30], i.e., the next generation of data*base* systems that seamlessly integrate structured data and unstructured data. The project will develop models that integrate unstructured data, structured data and annotations based on structured information retrieval [38], [42], [59].

3.2 Practical issues

The proposal presents novel ideas for developing a next-generation architecture for distributed information retrieval. Obviously, there are a number of practical issues that need to be resolved, and questions to be answered. Some of them are easily asked and easily answered:

Will local search providers ever participate in such a system? Yes, we think so. Companies want people to find their web pages: Companies and site owners spend a lot of money on *search engine positioning* [55] already. If a distributed system exists – of course there is a bootstrapping problem – it is likely that they would invest in getting more web traffic by providing local search. Note however, that the project does not depend on communities: The project will produce theory (retrieval models), technical solutions, and simulations of distributed information retrieval to test those solutions.

Are local sites fast enough to satisfy the latency requirements of a search engine? [54]. Today, local search on a significant number of sites might not be fast enough to answer a query in, say, 1 second. These systems will not be able to bid high, and therefore will not be selected unless there is no other system that produces quality results for the keywords. Furthermore, the efficiency of local site search will probably improve in the future.

Do we need to provide wrappers for thousands of local search systems? [73]. No, instead of accessing existing web forms, there needs to be some agreed upon application programming interface (API, Google provides an API for AdWords). Buntine et al. [17] give a comprehensive list of protocols and standards for information retrieval including standards that can be extended for bidding on keywords such as SRU/CQL [21].

3.3 Research approach

This section will briefly clarify the research methodology and experimental techniques. An important aim of the project is to develop new formal models and approaches to decentralized search. This work will be guided by applying formal models and theories developed for keyword auctions, on models and theories developed for language use (algorithms and methods motivated by game theory and statistical language models). All approaches will be implemented into research prototypes. These prototypes will be evaluated using simulations of highly distributed environments on one or two machines. Evaluations will use benchmark test collections consisting of real-world data and real user queries, as well as user relevance judgments (i.e., which documents are relevant for this query?). The information retrieval community is developing such benchmark collections as collaborative efforts in yearly evaluation workshops such as the Text Retrieval Conference (TREC) [76], the Cross-language Evaluation Forum (CLEF) [10] and the Initiative for the Evaluation of XML Retrieval (INEX) [47]. Participation in these workshops assures access to human subjects to test the system, which would be costly for individual research groups and hard to replicate by other researchers. Current evaluation tasks that are relevant for the proposal are the TREC terabyte task [22] and enterprise task [24].

4 Conclusion

Traditional centralized search engines have reached a limit in the amount of web pages they can crawl and the size and freshness of their indexes. Furthermore, they experience fundamental difficulties in indexing the deep web, and in supporting semi-structured data and annotations. Finally, the monopoly power of traditional centralized search engines threatens further developments in internet search, especially for small communities and minority languages. This research proposal has the potential to break the monopoly position of the major search engines by giving communities of users the power to collectively provide search services. In [3], a position paper written by many of the leading researchers in the field of information retrieval, this is put as follows: Developing *massively distributed* systems that *leverage world-wide structured and unstructured data* is considered to be the grand challenge in information retrieval. In the position paper *language models* are considered to play a major role in the development of the field. Advancements here are of great importance.

The project's aim is to distribute internet search functionality in such a way that communities of users and/or federations of small search systems provide search services in a collaborative way. Instead of getting all data to a centralized point and process queries centrally, as is done by today's search systems, the project will distribute queries over many small autonomous search systems and process them locally. Distributed information retrieval is a well researched sub area of information retrieval, but it has not resulted in practical solutions for large scale search problems because of high administration costs of setting up large numbers of installations and because it turns out to be hard in practice to direct queries to the appropriate local search systems. In this project we will research a radical new approach to distribute search: *distributed information retrieval by means of keyword auctions*. Keyword auctions like Google's AdWords give advertisers the opportunity to provide targeted advertisements by bidding on specific keywords. Analogous to these keyword auctions, local search systems will bid for keywords at a central broker. They "pay" by serving queries for the broker. The broker will send queries to those local search systems that optimize the overall effectiveness of the system, i.e., local search systems that are willing to serve many queries, but also are able to provide high quality results. The project will approach the problem from three different angles: 1) modeling the local search system, including models for automatic bidding and multi-word keywords; 2) modeling the search broker's optimization using the bids, the quality of the answers, and click-through rates; 3) integration of structured data typically available behind web forms of local search systems with text search. The approaches will be evaluated using prototype systems and simulations on benchmark test collections.

Acknowledgements

This project is funded by a Vidi grant of the Netherlands Organisation for Scientific Research (NWO).

References

[1] M. Abrol, N. Latarche, U. Mahadevan, J. Mao, R. Mukherjee, P. Raghavan, M. Tourn, J. Wang, and G. Zhang. Navigating large-scale semi-structured data in business portals. In *Proceedings of the 27th VLDB Conference*, pages 663–666, 2001

[2] R. Akavipat, L.S. Wu, F. Menczer, A.G. Maguitman. Emerging Semantic Communities in Peer Web Search. In *Proceedings of the Workshop on Information Retrieval in Peer-to-peer networks (P2PIR)*, 2006

[3] J. Allan, J. Aslam, N. Belkin, C. Buckley, J. Callan, W.B. Croft, S. Dumais, N. Fuhr, D. Harman, D.J. Harper, D. Hiemstra, T. Hofmann, E. Hovy, W. Kraaij, J. Lafferty, V. Lavrenko, D. Lewis, L. Liddy, R. Manmatha, A. McCallum, J. Ponte, J. Prager, D. Radev, P. Resnik, S.E. Robertson, R. Rosenfeld, S. Roukos, M. Sanderson, R. Schwartz, A. Singhal, A. Smeaton, H. Turtle, E. Voorhees, R. Weischedel, J. Xu, C. Zhai,. Challenges in Information Retrieval and Language Modeling *SIGIR Forum 37*(1), 2003

[4] 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 57*(3), pages 347–358, 2006

[5] R Baeza-Yates, B Ribeiro-Neto. Modern information retrieval. *Addison Wesley*, 1999

[6] R. Baeza-Yates, C. Castillo, F. Junqueira, V. Plachouras, F. Silvestri. Challenges on Distributed Information Retrieval. International Conference on Data Engeneering (ICDE), 2007

[7] M. Baillie, L. Azzopardi, F. Crestani. Adaptive query-based sampling for distributed collections. In *Proceedings of the Conference on String Processing and Information Retrieval*, pages 316–328, 2006

[8] W. Balke, W. Nejdl, W. Siberski, and U. Thaden. Progressive distributed top-k retrieval in peer-to-peer networks. In *Proceedings of the 21st International Conference on Data Engineering (ICDE),* 2005

[9] C. Baumgarten. A probabilistic model for distributed information retrieval. In *Proceedings of the ACM SIGIR Conference on Research and development in information retrieval*, 1997

[10] M. Braschler and C. Peters. Cross-Language Evaluation Forum: Objectives, Results, Achievements. *Information Retrieval 7*(1), pp.5–29, 2004

[11] M. Bender, S. Michel, P. Triantafillou, Ge. Weikum and C. Zimmer. Minerva: collaborative P2P search. In: *Proceedings of the 31st International Conference on Very Large Databases (VLDB)*, 2005

[12] M.K. Bergman. The deep web: Surfacing hidden value. *Journal of Electronic Publishing 7*(1), 2001.

[13] H.E. Blok, V. Mihajlovic, G. Ramirez, T. Westerveld, D. Hiemstra, and A.P. de Vries. The TIJAH XML Information Retrieval System. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006

[14] P. Boncz, T. Grust, M. van Keulen, S. Manegold, J. Rittinger, and J. Teubner. MonetDB/XQuery: A Fast XQuery Processor Powered by a Relational Engine. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2006

[15] N. Borch. Social peer-to-peer for social people. In *Proceedings of the International Conference on Internet Technologies and Applications*, 2005.

[16] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine, In Proceedings of the 7th WWW Conference pages, 107–117, 1998

[17] W. Buntine, M.P. Taylor, and F. Lagunas. Standards for Open Source Information Retrieval, In *Proceedings of the Open Source Information Retrieval Workshop (OSIR)*, 2006

[18] J. Callan, Distributed information retrieval. In: Advances in information retrieval, *Kluwer Academic Publishers*. pages 127–150, 2000

[19] J. Callan and M. Connell. Query-based sampling of text databases. *ACM Transactions of Information Systems 19*(2), pages 97–130, 2001

[20] J. Callan, Z. Lu, and W.B. Croft. Searching distributed collections with inference networks, In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–28, 1995

[21] S. Chumbe, R. MacLeod, P. Barker, M. Moffat, and R. Rist. Overcoming the obstacles of harvesting and searching digital repositories from federated searching toolkits, and embedding them in VLEs, In *Proceedings 2nd International Conference on Computer Science and Information Systems*, 2006.

[22] C.L.A. Clarke, F. Scholer, and I. Soboroff, The TREC 2005 Terabyte Track. In *Proceedings of the 14th Text Retrieval Conference (TREC)*, 2005

[23] N. Craswell, P. Bailey, and D. Hawking. Server selection on the World Wide Web, In *Proceedings of the ACM Conference on Digital Libraries,* pages 37–46, 2000

[24] N. Craswell, A.P. de Vries, and I. Soboroff. Overview of the TREC 2005 Enterprise Track. In *Proceedings of the 14th Text Retrieval Conference (TREC)*, 2005

[25] N. Craswell and M. Szummer. Random Walks on the Click Graph, In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 239–246, 2007

[26] W.B. Croft, J. Lafferty. Language Modeling for Information Retrieval. *Kluwer Academic Publishers*, 2003

[27] F.M. Cuenca-Acuna, C. Peery, R.P. Martin, and T.D. Nguyen. PlanetP: Using Gossiping to Build Content Addressable Peer-to-Peer Information Sharing Communities. In *Proceedings of the 12th International Symposium on High Performance Distributed Computing (HPDC)*, 2003

[28] B. Edelman, M. Ostrovsky, and M. Schwartz. Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. NBER working paper 11765, 2005

[29] A. Fast, D. Jensen, and B. N. Levine. Creating social networks to improve peer-to-peer networking. In *Proceedings of 11th ACM SIGKDD Conference*, 2005.

[30] M. Franklin, A. Halevy, and D. Maier. From databases to dataspaces: a new abstraction for information management. *ACM SIGMOD Record 34*(4):27–33, 2005

[31] J. Giles. Internet encyclopedias go head to head. *Nature* 438:900–901, 2005

[32] L. Gravano, P. Ipeirotis, M. Sahami. Qprober: A system for automatic classification of hidden-web databases. *ACM Transactions of Information Systems 21*(1), pages 1–41, 2003

[33] D. Hawking. Challenges in Enterprise Search. In *Proceedings of the 15th Australasian Database Conference*, 2004

[34] D. Hiemstra. A probabilistic justification for using tf.idf term weighting in information retrieval. *International Journal on Digital Libraries 3*(2), pp. 131-139, 2000

[35] D. Hiemstra. Using Language Models for Information Retrieval. Ph.D. Thesis, *Centre for Telematics and Information Technology*, *University of Twente*, 2001

[36] D. Hiemstra. Term-specific smoothing for the language modeling approach to information retrieval: the importance of a query term. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 35-41, 2002

[37] D. Hiemstra. Statistical Language Models for Intelligent XML Retrieval. In H. Blanken, T. Grabs, H.J. Schek, R. Schenkel and G. Weikum (eds.) Intelligent Search on XML. *Lecture Notes in Computer Science, Springer-Verlag*, pages 107–118, 2003

[38] D. Hiemstra and R. Baeza-Yates. Structured Text Retrieval Models, In *Encyclopedia of Database Systems, Springer* (to appear)

[39] D. Hiemstra and D. van Leeuwen. Creating an Information Retrieval test corpus for Dutch, In Mariët Theune, Anton Nijholt and Hendri Hondop (eds.) *Selected Papers of the 12th meeting of Computational Linguistics in the Netherlands (CLIN-12)*, *Editions Rudopi*, pages 133–147, 2002

[40] D. Hiemstra and W. Kraaij. A Language Modeling Approach to TREC. In E. Voorhees and D. Harman (eds.). TREC: Experiment and Evaluation in Information Retrieval, *MIT Press*, pages 373-395, 2005

[41] D. Hiemstra, S.E. Robertson and H. Zaragoza. Parsimonious Language Models for Information Retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185, 2004

[42] D. Hiemstra, H. Rode, R. van Os and J. Flokstra. PFTijah: text search in an XML database system. In *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR)*, pages 12–17, 2006

[43] J. Kamps. Experiments with Document and Query Representations for a Terabyte of Text. In *Proceedings of the 15th Text Retrieval Conference (TREC), 2006.*

[44] I.A. Klampanos and J.M. Jose. An architecture for information retrieval over semi-collaborating Peer-to-Peer networks. *Proceedings of the 2004 ACM symposium on Applied computing*, pages 1078–1083. 2004

[45] W. Kraaij, T. Westerveld and D. Hiemstra. The Importance of Prior Probabilities for Entry Page Search. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pages 27–34, 2002

[46] S. Lahaie, D. Pennock, A. Saberi, and R. Vohra. Sponsored Search. In N. Nisan et al. (eds.) *Algorithmic Game Theory*, *Cambridge University Press*, 2007

[47] M. Lalmas and G. Kazai. Report on the ad-hoc track of the INEX 2005 workshop. *SIGIR Forum 40*(1):49–57, 2006

[48] S. Lawrence and C.L. Giles. Accessibility of Information on the Web, *Nature 400*(107), 1999

[49] J. Li, T. Loo, J. Hellerstein, F. Kaashoek, D. Karger, and R. Morris. On the Feasibility of Peer-to-Peer Web Indexing and Search. In *Proceedings of the 2nd International Workshop on peer-to-peer systems (IPTPS)*, pages 207–215, 2003

[50] J. List, V. Mihajlovic, G. Ramirez, A.P. de Vries, D. Hiemstra, and H.E. Blok. Tijah: Embracing IR Methods in XML Databases. *Information Retrieval Journal 8*(4), pp. 547–570, *Kluwer Academic Publishers*, 2005

[51] J. Lu and J. Callan. User modeling for full-text federated search in peer-to-peer networks. In *Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006

[52] J. Lu and J. Callan. Full-text federated search of text-based digital libraries in peer-to-peer networks. *Journal of Information Retrieval 9*(4), 2006

[53] T. Luu, F. Klemm, I. Podnar, M. Rajaman, K. Aberer. ALVIS Peers: A Scalable Full-text Peer-toPeer Retrieval Engine. In: *Proceedings of the Workshop on Information Retrieval in Peer-to-Peer Networks (P2PIR)*, 2006

[54] J. Madhavan, A. Halevy, S. Cohen, X. Dong, S.R. Jeffrey, D. Ko, and C. Yu. Structured Data Meets the Web: A Few Observations. *IEEE Computer Society Technical Committee on Data Engineering*, 2006

[55] F.W. Marckini. Search Engine Positioning, *Wordware Publishing*, 2001

[56] A. Mehta, A. Saberi, U. Vazirani, and V. Vazirani. AdWords and Generalized Online Matching. *Journal of the ACM 54*(5), 2007

[57] M. Melucci and A. Poggiani. A study of a weighting scheme for information retrieval in hierachical peer-to-peer networks, In *Proceedings of the European Conference on Information Retrieval. Springer*, pages 136–147, 2007

[58] S. Michel, M. Bender, P. Triantafillou and G. Weikum. IQN Routing: Integrating Quality and Novelty in P2P Querying and Ranking. In: *Proceedings of the 10th International Conference on Extending Database Technology (EDBT)*, 2006

[59] V. Mihajlovic, H.E. Blok, D. Hiemstra and P.M.G. Apers. Score Region Algebra: Building a Transparent XML-IR Database. In *Proceedings of the fourteenth International Conference on Information and Knowledge Management (CIKM)*, pages 12–19, 2005

[60] H. Nottelman and N. Fuhr. A decision-theoretic model for decentralized query routing in hierarchical peer-to-peer networks, In *Proceedings of the European Conference on Information Retrieval. Springer*, pages 148–159, 2007

[61] T. O'Reilly, What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software. *O'Reilly Media*, September 2005

[62] G. Paltoglou, M. Salampasis and M. Satratzemi. Result merging algorithm using multiple regression models. In *Proceedings of the European Conference on Information Retrieval. Springer*, pages 173–184, 2007

[63] I. Podnar, T. Luu, M. Rajman, F. Klemm, K. Aberer, "A Peer-to-Peer Architecture for Information Retrieval Across Digital Library Collections", *Proceedings of the European conference on research and advanced technology for digital libraries (ECDL)*, September 2006

[64] J.A. Pouwelse, P. Garbacki, J. Wang, A. Bakker, J. Yang, A. Iosup, D.H.J. Epema, M. Reinders, M. van Steen, H.J. Sips. Tribler: A social-based Peer-to-Peer system. In *Proceedings of the 5th International Workshop on Peer-to-Peer Systems* (IPTPS), 2006

[65] P. Reynolds and A. Vahdat. Efficient peer-to-peer keyword searching. In *Proceedings of the ACM Middleware Conference*, pages 21–40, 2003.

[66] M. Shokouhi. Segmentation of Search Engine Results for Effective Data-Fusion. In Proceedings of the European Conference on Information Retrieval (ECIR), pages 185–197, 2007

[67] M. Shokouhi, M. Baille, and L. Azzopardi. Updating collection representations for federated search. In *Proceedings of the 30th International ACM SIGIR on Information Retrieval*, pages 511–518, 2007

[68] M. Shokouhi, J. Zobel, S. Tahaghoghi, and F. Scholer. Using query logs to establish vocabularies in distributed information retrieval, *Information Processing and Management 43*, pages 169–180, 2007

[69] L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In *Proceedings of the 26th International ACM SIGIR on Information Retrieval*, pages 298–305, 2003

[70] L. Si and J. Callan. A semi-supervised learning method to merge search engine results. *ACM Transactions on Information Systems 21*(4), pages 457–491, 2003

[71] T. Suel, C. Mathur, J.W. Wu, J. Zhang, A. Delis, M. Kharrazi, X. Long, and K. Shanmugasundaram. Odissea: A Peer-to-Peer Architecture for Scalable Web Search and Information Retrieval. *Proceedings of the International workshop on the Web and Databases*, 2003.

[72] C. Tang, Z. Xu, and M. Mahalingam. pSearch: information retrieval in structured overlays. In *Proceedings of ACM SIGCOMM Computer Communication Review*, 2003

[73] P. Thomas and D. Hawking. Evaluating sampling methods for uncooperative collections, In *Proceedings of the 30th International ACM SIGIR on Information Retrieval*, pages 503–510, 2007

[74] K.J. Tinselboer. The use of rare key indexing for distributed web search, *Master's Thesis, University of Twente*, September 2007

[75] H.R. Varian. Position auctions. *International Journal of Idustrial Organization*, 2006

[76] E. Voorhees and D. Harman (eds.). TREC: Experiment and Evaluation in Information Retrieval. *MIT Press*, 2005

[77] S. Waterhouse, D.M. Doolin, G. Kan, and Y. Faybishenko. Distributed Search in P2P Networks. *IEEE Internet Computing 6*(1), pages 68-72, 2002

[78] T. Westerveld, W. Kraaij and D. Hiemstra. Retrieving Web Pages using Content, Links, URLs and Anchors. In *Proceedings of the 10th Text Retrieval Conference*, pages 663-672, 2002

[79] J. Xu and W.B. Croft. Cluster-based language models for distributed retrieval, In *Proceedings of the 21st International ACM SIGIR on Information Retrieval*, pages 254–261, 1999

[80] Y. Yang, R. Dunlap, M. Rexroad, and B.F. Cooper. Performance of Full Text Search in Structured and Unstructured Peer-to-Peer Systems. In: *Proceedings of the The 25th Conference on Computer Communications (IEEE INFOCOM)*, 2006

[81] R. Zagers. A decision system for Google Adwords' (in Dutch), *Master's Thesis*, University of Twente, done at Gladior, Enschede, 9 February 2007

[82] J. Zhang, and T. Suel. Efficient query evaluation on large textual collections in a peer-to-peer environment. In: *Proceedings of the Fifth IEEE International Conference on Peer-to-Peer Computing* (P2P'05), 2005

[83] Attack of the Eurogoogle. *The Economist Technology Quarterly*, March 2006, pp.8-9.