Open Competitie 2005
Exacte Wetenschappen

---

1 Project Details

1.a) Project Title. EFfective FOcused Retrieval Techniques.

1.b) Project Acronym. EFFORT.

2 Project Content

2.a) Summary. The advent of the Internet marked the beginning of the information society, by giving us access to an unprecedented amount of data and information. But as the Internet is growing, more and more information is becoming out of reach of large-scale, general purpose web search engines. There are alternatives such as web directories and specialized search engines targeting a specific domain, as well as envisaged alternatives like the semantic web. Each alternative has its own representation of web data and each has its own strengths and weaknesses. Combining these representations in a single framework has the potential to provide very accurate and focused search of web data.

This research proposal will develop an approach for combining multiple representations of web information in a common framework based on statistical language models. In this framework, it will be possible, for example, to derive models of the actual language-use of web pages to distinguish between arts, business, entertainment, education, etc. Similarly, it will be possible to derive models of the structure of web pages to distinguish between blogs, FAQs, personal web pages, cultural heritage pages, etc. The envisaged techniques have to be robust to all kinds of errors, ranging from imperfect information extraction techniques to imprecise queries formulated by the average web search engine user. An important aspect of any new technique in a web-setting is that they have to scale up to terabyte-sized collections. We plan to develop so-called *parsimonious* models to derive compact representation and to handle dependencies between representations of the data.

2.b) Abstract for laymen (in Dutch). Information Retrieval is het vakgebied dat zich bezig houdt met de ontsluiting van grote documentencollecties. Voor het grote publiek is dit vrijwel synoniem met het concept van Internetzoekmachines zoals Google. Bedrijven als Google hadden tot voor kort voldoende aan standaard tekstzoekfaciliteiten gecombineerd met een analyse van de hyperlink structuur van het web. Echter, internetzoekapplicaties worden steeds complexer. In de eerste plaats door het belang van handmatige metadata van

bijvoorbeeld het Open Directory Project en mogelijk in de toekomst het Semantic Web; en in de tweede plaats door de opkomst van gespecialiseerde zoekmachines, voor internetwinkelen, voor nieuws, of voor wetenschappelijke documenten. Deze ontwikkelingen leveren gestructureerde informatie op in de vorm van metadata-classificaties in het geval van het Semantic Web, en in de vorm van geëxtraheerde data zoals productnaam en prijs in het geval van gespecialiseerde zoekmachines. Dit projectvoorstel heeft als doel een raamwerk te ontwikkelen voor het beheren en combineren van zulke gestructureerde informatie. Zo'n raamwerk zal zeer accurate en preciese zoekfuncties mogelijk maken.

Het voorgestelde raamwerk, dat gebruik gaat maken van zogenaamde statistische taalmodellen, zal het mogelijk maken om onderscheid te maken tussen bijvoorbeeld recreatie, kunst en zakelijk, enz. op grond van het taalgebruik in webpagina's, en onderscheid tussen blogs, FAQs, persoonlijke webpagina's, enz. op grond van de opmaakconventies die een pagina gebruikt. Een belangrijk aspect van de voorgestelde aanpak is de toepassing van zuinige ("parsimonious") taalmodellen die meerdere documentrepresentaties combineren tot een compacte representatie, daarbij rekening houdend met afhankelijkheden tussen de representaties.

3 Description of the Proposed Research

3.1 Scientific Problem and Expected Results

The Internet gives us access to an unprecedented amount of data and information [71]. This creates great challenges and opportunities. As to challenges, the growth of information prompts the need for methods that can help to organize, cluster and classify information, and improve ways of accessing it. This has led to a resurgence of interest in the field of information retrieval [5, 13, 39, 83]. Large-scale, general purpose web search engines (most notably Google) have been quite successful in keeping up with the size of the web by adding new sources of information to traditional full-text indexes: for instance anchor texts and hyperlink structure. However, the web is growing out of reach of these techniques, and companies like Google have started to offer specialized web search services focusing on for instance internet shopping (Froogle) and scientific documents (Google Scholar inspired by CiteSeer [15]). Such services use lay-out analysis techniques and simple information extraction techniques that exploit web conventions—extracting e.g., product names, prices, author names, etc.—to come up with domain-dependent structured document representations that are far more complex than the full-text/anchor text/link structure representations. As such, those services provide very accurate and focused search.

However, specialized search engines only provide focused search if the user is *first* able to find the specialized search engine of his/her choice, if one that caters for the user's problem exists at all. Whereas specialized search engines

are part of the answer to the increasing size of the web—they identify more structured information and provide more focused search—they also introduce new information overload problems. We believe we can have the best from both world: very focused and accurate search without the need for the user to preselect the domain beforehand. The main *research problem* of the EfFoRT proposal is the following:

- Can we provide accurate and focused search in a large-scale general purpose web search scenario by adding structure and combining multiple, complex representations within a common information retrieval *modeling* framework?

There is an increasing amount of structure in documents on the web. Like the specialized search engines mentioned above, we intend to develop structured document representations, where the structure is derived from the document text, the document structure, the URL, the hyperlink structure, time, geographic location, text classification, manually assigned metadata, and more. All these sources of evidence can provide crucial retrieval cues. However, at the same time these sources may be interdependent in various complex ways. That is, when we combine two features that are effective in isolation, this can easily result in poor performance due to overestimation. The increasing amount of structure on the web creates a need for robust methods that can exploit these additional retrieval cues. Against this background, the EfFoRT proposal aims to deliver the following results:

1. To model dependencies and remove redundancy between representations. In our opinion, current attempts to model structurally complex document representations suffer from a fundamental problem: they fail to adequately handle dependencies between different modeling hierarchies. It has been shown that state-of-the-art document representation approaches that combine evidence from multiple sources overestimate words that are too general (such as function words, which are redundant because they are modeled well by general models of English) unless special care is taken [86, 87]. Properly handling dependencies will also help identifying words that seem to contribute significantly to only one of the representations but that are unlikely to contribute to retrieval performance, like spelling mistakes.

2. To model and exploit the hierarchical structure of the internet. Currently, web search is almost always web *page* search, but pages are part of a web site, web sites are organized by domains, and domains are organized by top-level domains. Top-level domains provide a crude but potentially powerful way to cluster web information by topic (.com, .edu, .org, etc.) and by geographical location (using country top-level domains like .nl and .uk). The hierarchical web structure can be used to improve link-based methods [85], but also to provide

focus: a 'relevant' page may be more interesting if it is part of a site dedicated to the problem at hand. Approaches that try to exploit hierarchical structure to improve information retrieval effectiveness are currently being developed by researchers working on full-text search in XML data [3, 24, 25, 57]. However, the dependency problem is especially apparent in these approaches, because a single word occurrence (at least in theory) retrieves as many elements (from a single XML document) as its depth in the XML tree. That is, the occurrence of a word in a paragraph, would return the subsection that contains the paragraph, the section that contains the subsection, the chapter that contains the section, etc. Focused retrieval approaches are needed here.

3. To model and exploit topical language similarity. We will investigate the particular language use for specific types of web pages, and build localized models for them. For example, we can consider the type of language used on home pages. As a next step, we can differentiate between (home) pages of people and those of organizations; etcetera. Another example concerns pages about products in on-line shops, which again can be further broken down into particular product categories, etc. Note that this need not be restricted to factual topics: we can model the opinionated language usage on product reviews or blogs. We can anchor the topical content types on Internet directories as provided by the Open Directory, Yahoo!, Google, or on-line encyclopedias such as Wikipedia [1, 14, 66, 78]. Particular language usage provides a layer of *topical language models* that can be incorporated in the language modeling framework.

4. To model and exploit topical structure similarity. Whereas a typical search engine will disregard almost all of the document markup, and focus primarily on the textual content presented to the user, we expect that there are valuable retrieval cues in structural aspects of web pages. This is highly related to the emergence of web conventions, in which specific types of web content have adopted a similar look-and-feel. As such there is great structural similarity between, for example, home pages of people, on-line product pages, FAQs, blogs, etc. Abstracting the structure of the pages provides another layer of *structural language models* that can be incorporated in the modeling framework. This strategy is well known in computational linguistics, in particular in data-oriented parsing [9]. Specifically, we will adopt a similar approach to overcoming data-sparseness by breaking the structural trees in small fragments. We will consider both tag-name labeled and unlabeled trees of the document structure.

5. To model and exploit the user's context. The bottle-neck for providing more focused retrieval is in the shallowness on the client-side, i.e., users who provide no more than a few keywords to express their complex information needs. An approach would be to let users articulate their search request in a structured query language [45]. Here, our main approach is to try to avoid that

the user has to provide explicit references to the derived metadata and added structure, at least not in an initial query. We aim to use the initially returned set of documents, and to use the user's context implicitly and explicitly. Implicit information about the users geographical might for instance come from his/her IP number, or implicit user preferences might be derived from click-through information. Explicit information might come from explicit interaction similar to Google's spelling suggestions, for instance "Do you want to focus on *sports*?", "Do you want to focus on documents from around 11 September 2001?", "Do you want to focus on documents from Europe?", "Do you need an FAQ?", "Are you looking for a person's home page?", or "Do you want to focus on products?" [52, 66].

3.2 Research Method

Our modeling framework will be based in so-called *statistical language models* for information retrieval [29]. Language models for retrieval are easily combined with information from other models. For instance, language models have been successfully combined with statistical translation models [7] and as such have been applied to cross-language retrieval [23, 31, 84]. Our approach is to go beyond the standard "document as a bag-of-words" models by bringing more and more sources of evidence into the models. We have successfully combined the basic retrieval model with models of non-content information that use for instance link information and URL-type in web retrieval [33, 42, 46, 48]. We will extend existing language modeling approaches in several ways.

Relevance models Our approach to implicit feedback is related to relevance models [53, 54] in which the set of initially retrieved documents is included in the model as a layer between the document and the collection model. For example, we believe that there is great potential in the combination of metadata (either in the form of e.g., the Yahoo! directory, or metadata from the document itself) with derived representations. Using such metadata, we will be able to derive topical models, i.e., models of the language typically used in documents on a certain topic. For instance, a model built from documents in the Yahoo! category "Science $>$ Computer Science" can be used to provide more targeted search, searching e.g., "Java" in that context would exclude the Indonesian island [66]. Topical language models can also be used to classify (initially retrieved) documents using text categorization techniques [70], similar to [12]. This, in turn, provides crucial information for adjusting smoothing parameters, for result clustering, or for asking follow-up questions like "do you want to focus on sports?"

Parsimonious language models Usually, each language model representation is defined independently from the others and they are combined later on. However, independent definitions of language modeling components lead to large,

redundant combined representations. To model structurally complex document representations, we need a method that rewards *parsimony*. Parsimonious models [75] explicitly address the relation between several representations of the document. As such, they effectively model for each partial representation what it adds to the document representation as a whole, thus avoiding redundancy. We, and others, have recently shown that parsimonious models improve retrieval performance in both text search [35] and image search [82]. Techniques from parsimonious language models allow us to define those representations in a layered fashion by explicitly addressing dependencies between document representations. As additional but important bonus, avoiding redundancy results in significantly *smaller* models. This leads to a reduction of storage space, and a reduction of query processing time—two key requirements for effective large-scale web retrieval.

Region models Region models were developed for structured document retrieval [4, 11, 16, 20, 38, 67]. They provide a well-defined behavior as well as a simple query language that allows for simple structured region queries like: "give me web pages containing X, contained by web sites containing Y". Recently [34], we have shown a remarkable one-to-one relation between region queries and language modeling approaches: A wide variety of language modeling approaches—approaches that *seem* to be unrelated to structured document retrieval, such as simple ad-hoc search, web search, cross-language retrieval, and video retrieval—can be expressed as region queries. This offers an enormous flexibility in terms of the modeling approaches without the need to re-index the collection. Moreover, we conjecture that region models will give additional theoretical insights in how simple "bag-of-words models" can be enhanced by structure [56].

Theory guided by experimentation The history of information retrieval is a showcase of theoretical progress going hand-in-hand with experimental evaluation. The scientific evaluation of information retrieval systems is rooted in the Cranfield experiments [18, 19]. This has been continued in recent years within the framework of the Text REtrieval Conference [TREC, 26, 81], and its various regional and task-specific counterparts such as CLEF [17], NTCIR [61], and INEX [37].

EFFORT will re-use existing collections and test-suites as far as possible and extend them were the need arises. In particular, we are considering the following collections: First, the `W3C` corpus of TREC's Enterprise track. Second, the `.GOV2` corpus of TREC's Terabyte track. Third, the EUROGOV corpus of CLEF's WebCLEF track. Fourth, where the need arises, we will supplement these with new corpora. As to the last point, we intend to do focused crawling of specialized sub collections, such as pages of products [13, 14]. We have built up substantial experience with focused crawling during the construction of

the EUROGOV corpus, a heterogeneous collection of European governmental information [72, 73]. Finally, we intend to organize a special focused retrieval track within one of the evaluation campaigns.

3.3 Scientific Significance

The current rate of growth of information on the Internet leads to ineffectiveness and myopia, unless it is paralleled with a similar growth in effectiveness of information retrieval techniques. The wealth of information available on the Internet creates an urgent need for methods that can help organize, cluster and classify information, and improve ways of accessing it (such as information disclosure and knowledge distillery). Internet search engines, such as Google, give millions of users per day access to over 8 billion web pages. Given a user's information need, search engines return a ranked list of relevant documents. Yet most of the queries return thousands of pages, of which seldom more than a handful is actually inspected. This is a tell-tale sign of the need to revisit the classic use of relevance as the fundamental notion in the information sciences [69]. Other examples are information pinpointing approaches such as XML retrieval and question answering. In XML retrieval [44], individual XML elements are the unit of retrieval; they are judged on both a relevance-dimension and a coverage-dimension (the extent to which non-relevant information is present). In question answering [80] the task is to return exact answers to questions, rather than complete documents potentially containing the answer. In web retrieval, it has been suggested to supplement relevance with a notion of importance or authority [10, 47]. The effectiveness of PageRank [10] clearly shows the potential of techniques making sense of the noisy structure of the web.

The current proposal significantly extends the types of information that are taken into account by today's search engines. It will introduce techniques for exploiting topical language usage, and for topical structure similarity. These powerful techniques will provide more focused retrieval, by going beyond the one-size-fits-all approach of today's general purpose web search engines. As such, it holds the promise to significantly contribute to the next generation of focused web search engines. The importance becomes even more evident when considering that the average web user refuses to articulate their complex information needs in more than 2-3 words [40, 76]. The onus of effective retrieval based on such a highly ambiguous search statement is placed entirely on the search engine. In our opinion, the proposed language modeling techniques can make significant advances here.

In web retrieval there is a longstanding debate on the effectiveness of link-based methods [27]. In our opinion, the complex dependencies between different sources of web evidence is one of the main reasons for the diverging results. Careful modeling of the dependencies between various sources of evidence will allow us to isolate their unique contributions. This seems to be a crucial step towards furthering our general understanding of the precise role of various sources

of evidence. We firmly believe that significant progress in focused web retrieval can only be achieved by a mixture of theoretical and experimental work, by building models that closely correspond to theoretical intuitions, and by interpreting experimental results in terms of such transparent models.

3.4 Related Research

The EfFoRT proposal addresses robust and focused retrieval techniques that deal with the increasing amount of structure on the web. Whereas part of this is orchestrated, for example by the adoption of Semantic Web techniques [8], much of it is due to the self-organizing nature of the Web [6, 50]. As a result, the structure on the web is rich but highly heterogeneous and noisy, and there is a need for techniques that help organize noisy structured information. Important initial steps have been taken here. Craswell et al. [21] highlighted the importance of compact document representations based on anchor-texts for the specialized task of home page finding. For the same home page finding task, Kraaij et al. [48] showed the importance of URL information, and introduced mixture language models that incorporate anchor text and URL information. This approach was extend by Ogilvie and Callan [62], addressing now more general known-item searches. Finally, using similar approaches, Kamps et al. [42, 46] addressed various URL and link priors for both known-item search and topic distillation. Language models have been used for structured data in several other ways. Ogilvie and Callan [63] applies techniques from stochastic context-free parsing to XML retrieval. In related work outside the language modeling framework, Craswell et al. [22] address a range of query-independence evidence for BM25. They propose transformations that take into account dependencies between the feature values and the content-scores. These methods nicely complement our current work [35] in which we focused, so far, on dependencies with the query-dependent evidence.

The EfFoRT proposal will substantially extend on earlier approaches, by taking into account topical language usage, structural similarity, hierarchical web structure, and the user's context. EfFoRT will develop a generic approach for combining multiple, complex representations of heterogeneous web information in a common framework based on statistical language modeling.

The term *language models* originates from probabilistic models of language generation developed for automatic speech recognition systems in the early 1980s (see e.g., [65]), but at that time they had already been in existence for about 70 years. Language models have had quite an impact on information retrieval research. Within three years after their first application to information retrieval in 1998 by Ponte and Croft [64], Hiemstra and Kraaij [28, 32] and Miller et al. [58, 59], the 2001 ACM Conference on Research and Development in Information Retrieval had two sessions on the use of language models [49]. The recent road map for future information retrieval research: "Challenges in Information Retrieval and Language Modeling" [2] identifies contextual retrieval

and global information access as particularly important long-term challenges.

Essential in these approaches is so-called *smoothing*. Smoothing is the task of re-evaluating the probabilities, in order to assign some non-zero probability to query terms that do not occur in a document. As such, smoothed probability estimation is an alternative for maximum likelihood estimation. The standard language modeling approach to information retrieval uses a linear interpolation smoothing of the document model with a general collection model [7, 32, 51, 59, 60, 74]. There are other approaches to smoothing language models (see e.g., [41]), some of which have been suggested for information retrieval as well, for instance smoothing using the geometric mean and backing-off by Ponte and Croft [64]; and Dirichlet smoothing and absolute discounting suggested in a study by Zhai and Lafferty [86]. In a sense, all smoothing approaches somehow combine two representations of the data (a document model and a collection model) into a new representation. We plan to use similar approaches to combine many more document representations into a single representation. A major problem with such combined representation is that they contain a lot of redundancy. This problem can already be observed for the standard linear interpolation smoothing model: Words that are most frequent in the document model (like "the", "a" and "it") are already very well explained by the collection model. Such words are often treated as *stop words*, they are ignored because they take a lot of system resources, while they are not very likely to contribute to search performance [30]. Currently, systems use manually constructed stop lists or some other ad-hoc considerations to ignore stop words. When combining several document representations in a multiple level language model, we need a more principled way to ignore words. We plan to build models in a layered fashion where each level models the *difference* with the more general models. In a multiple level language model, some levels might be defined as models of specific topics as done by Hofmann [36]. As such, a unigram model of general language, modified by a small number of topic-specific unigram models, each of which has parameters for only a small number of topic-specific specialist terms and phrases, would be a very *parsimonious* and efficient model for a collection of documents [75].

The idea of *parsimonious* language models bears some resemblance with early work on information retrieval by Luhn [55] who specifies two frequency cut-off lists, an upper and a lower, to exclude non-significant words. The words exceeding the upper cut-off are considered to be common and those below the cut-off rare, and therefore not contributing to the content of the document [79]. Simple frequency cut-offs are not applicable for the complex document representations we envisage, for which each partial representation only models those aspects that are not covered by the other representations. Similar approaches have been applied to back-off language models for automatic speech recognition: Stolcke [77] and Sankar et al. [68] describe approaches to reduce the language model size by pruning n-grams based on relative entropy between the original models and the pruned model.

Region models were developed for structured document retrieval [4, 11, 16, 20, 38, 67], but have always been a bit of an outsider in information retrieval modeling because, unlike other retrieval models, they do not focus on the ranking of retrieved results. Baeza-Yates and Ribeiro-Neto [5] put it as follows in their textbook on Modern Information Retrieval: *"In fact, this* (appropriate ranking for region models) *is an actual, interesting, and open research problem"*. We have recently made practical advancements in this area [56] as well as theoretical discoveries on connecting region modeling approaches to the language modeling approach [34]. Parsimonious relevance models and region models with appropriate ranking both have the potential of causing a break-through in information retrieval theory similar to language models in 1998. We believe these approaches will be very effective in providing accurate and focused search by adding structure to, and combining representations of, web data. Research in these directions is urgent.

4 Literature

[1] D. Ahn, V. Jijkoun, J. Kamps, G. Mishne, K. Müller, M. de Rijke, and S. Schlobach. The University of Amsterdam at TREC 2004. In *TREC 2004 Working Notes*, pages 43–56. National Institute for Standards and Technology, 2004.

[2] J. Allan, J. Aslam, N. Belkin, C. Buckley, J. Callan, W. Croft, S. Dumais, N. Fuhr, D. Harman, D. 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. Robertson, R. Rosenfeld, S. Roukos, M. Sanderson, R. Schwartz, A. Singhal, A. Smeaton, H. Turtle, E. Voorhees, R. Weischedel, J. Xu, and C. Zhai. Challenges in information retrieval and language modeling. *SIGIR Forum*, 37(1), 2003.

[3] S. Amer-Yahia, C. Botev, and J. Shanmugasundaram. TeXQuery: A full-text search extension to XQuery. In *Proceedings of the 13th conference on World Wide Web*, pages 583–594, 2004.

[4] R. Baeza-Yates and G. Navarro. Proximal nodes: A model to query document databases by content and s tructure. *ACM Transactions on Information Systems*, 15(4):401–435, 1997.

[5] R. Baeza-Yates and B. Ribeiro-Neto, editors. *Modern Information Retrieval*, 1999. ACM Press, New York and Addison Wesley Longman, Harlow.

[6] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[7] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd ACM Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 222–229, 1999.

[8] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(3):34–43, 2001.

[9] R. Bod, R. Scha, and K. Sima'an, editors. *Data-Oriented Parsing*. CSLI Publications, 2004.

[10] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine.

In *Proceedings of the 7th International World Wide Web Conference*, pages 107–117. Elsevier Science, New York, 1998.

[11] F. J. Burkowski. Retrieval activities in a database consisting of heterogeneous collections of structured text. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '92)*, pages 112–125, New York, NY, USA, 1992. ACM Press.

[12] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

[13] S. Chakrabarti. *Mining the Web: Analysis of Hypertext and Semi Structured Data*. Morgan Kaufmann, 2002.

[14] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: A new approach to topic-specific Web resource discovery. *Computer Networks*, 31:1623–1640, 1999.

[15] CiteSeer. CiteSeer: Scientific literature digital library, 2005. `http://citeseer.ist.psu.edu/`.

[16] C. L. A. Clarke, G. V. Cormack, and F. J. Burkowski. An algebra for structured text and a framework for its implementation. *The Computer Journal*, 38:43–56, 1995.

[17] CLEF. Cross Language Evaluation Forum, 2005. `http://www.clef-campaign.org/`.

[18] C. W. Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, College of Aeronautics, Cranfield UK, 1962.

[19] C. W. Cleverdon. The Cranfield tests on index language devices. *Aslib*, 19:173–192, 1967.

[20] M. Consens and T. Milo. Algebras for querying text regions. In *Proceedings of the ACM Conference on Principles of Distribute d Systems*, pages 11–22, 1995.

[21] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In Kraft et al. [49], pages 250–257.

[22] N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05)*, pages 416–423, New York, NY, USA, 2005. ACM Press.

[23] M. Federico and N. Bertoldi. Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th ACM Conference on Research and Development in Information Retrieval (SIGIR'02)*, pages 167 – 174, 2002.

[24] N. Fuhr and K. Großjohann. XIRQL: A query language for information retrieval in XML documents. In Kraft et al. [49], pages 172–180.

[25] N. Fuhr and K. Großjohann. XIRQL: An XML query language based on information retrieval concepts. *ACM Transactions on Information Systems*, 22:313–356, 2004.

[26] D. K. Harman, editor. *The First Text REtrieval Conference (TREC-1)*, 1993. National Institute for Standards and Technology. NIST Special Publication 500-207.

[27] D. Hawking and N. Craswell. Very large scale retrieval and web search. In E. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[28] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, pages 569–584, 1998.

[29] D. Hiemstra. *Using language models for information retrieval*. PhD thesis, Centre for Telematics and Information Technology, University of Twente, 2001.

[30] D. Hiemstra. Term-specific smoothing for the language modeling approach to information retrieval: The importance of a query term. In *Proceedings of the 25th ACM Conference on Research and Development in Information Retrieval (SIGIR'02)*, pages 35–41, 2002.

[31] D. Hiemstra and F. de Jong. Disambiguation strategies for cross-language information retrieval. In *Proceedings of the third European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, pages 274–293, 1999.

[32] D. Hiemstra and W. Kraaij. Twenty-One at TREC-7: Ad-hoc and cross-language track. In *Proceedings of the seventh Text Retrieval Conference TREC-7*, pages 227–238. NIST Special Publication 500-242, 1998.

[33] D. Hiemstra and W. Kraaij. A language modeling approach to TREC. In E. Voorhees and D. Harman, editors, *TREC: Experimentation and Evaluation in Information Retrieval*. MIT Press, 2005.

[34] D. Hiemstra and V. Michajlovic. A database approach to information retrieval: The remarkable relationship between language models and region models. Technical Report 05-35, Centre for Telematics and Information Technology, 2005.

[35] D. Hiemstra, S. 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. ACM Press, New York NY, 2004.

[36] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd ACM Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 50–57, 1999.

[37] INEX. INitiative for the Evaluation of XML retrieval, 2005. `http://inex.is.informatik.uni-duisburg.de:2003/`.

[38] J. Jaakkola and P. Kilpelainen. Nested text-region algebra. Technical Report CR-1999-2, Department of Computer Science, University of Helsinki, 1999.

[39] P. Jackson and I. Moulinier. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*, volume 5 of *Natural Language Processing*. John Benjamins Publishing Company, Amsterdam, 2002.

[40] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36:205–339, 2000.

[41] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1997.

[42] J. Kamps. Web-centric language models. In *Proceedings of the Fourteenth ACM Conference on Information and Knowledge Management (CIKM 2005)*. ACM Press, New York NY, USA, 2005.

[43] J. Kamps, M. de Rijke, and B. Sigurbjörnsson. Length normalization in XML retrieval. In M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, editors, *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 80–87. ACM Press, New York NY, USA, 2004.

[44] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. XML retrieval: What to retrieve? In C. Clarke, G. Cormack, J. Callan, D. Hawking, and A. Smeaton, editors, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 409–410. ACM Press, New York NY, 2003.

[45] J. Kamps, M. Marx, B. Sigurbjörnsson, and M. de Rijke. Structured queries in XML retrieval. In *Proceedings of the Fourteenth ACM Conference on Information and Knowledge Management (CIKM 2005)*. ACM Press, New York NY, USA, 2005.

[46] J. Kamps, G. Mishne, and M. de Rijke. Language models for searching in Web corpora. In E. M. Voorhees and L. P. Buckland, editors, *The Thirteenth Text REtrieval Conference (TREC 2004)*. National Institute of Standards and Technology. NIST Special Publication 500-261, 2005.

[47] J. M. Kleinberg. Authoritative structures in a hyperlinked environment. *Journal of the ACM*, 46:604–632, 1999.

[48] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In K. Järvelin, M. Beaulieu, R. Baeza-Yates, and S. H. Myaeng, editors, *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34. ACM Press, New York NY, USA, 2002.

[49] D. H. Kraft, W. B. Croft, D. J. Harper, and J. Zobel, editors. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001. ACM Press, New York NY, USA.

[50] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *Proceedings of the 8th International World-wide web conference WWW8*, pages 403–415. Elsevier Science, Amsterdam, 1999.

[51] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization. In *Proceedings of the 24th ACM Conference on Research and Development in Information Retrieval (SIGIR'01)*, pages 111–119, 2001.

[52] B. Larssen, editor. *ACM SIGIR 2004 workshop on Information Retrieval in Context*, 2004.

[53] V. Lavrenko and W. Croft. Relevance-based language models. In *Proceedings of the 24th ACM Conference on Research and Development in Information Retrieval (SIGIR'01)*, pages 120–128, 2001.

[54] V. Lavrenko and W. Croft. Relevance models in information retrieval. In W. Croft and J. Lafferty, editors, *Language Modeling for Information Retrieval*, pages 11–56. Kluwer Academic Publishers, 2003.

[55] H. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159–165, 1958.

[56] V. Michajlovic, H. E. Blok, D. Hiemstra, and P. M. G. Apers. Score region algebra: Building a transparent XML-IR database. In *Proceedings of the Fourteenth ACM Conference on Information and Knowledge Management (CIKM 2005)*. ACM Press, New York NY, USA, 2005.

[57] V. Mihajlovic, G. Ramirez, A. de Vries, D. Hiemstra, and H. Blok. Tijah at inex 2004: Modeling phrases and relevance feedback. In *Proceedings of the 3rd Initiative on the Evaluation of XML Retrieval (INEX 2004)*, 2004.

[58] D. Miller, T. Leek, and R. Schwartz. BBN at TREC-7: using hidden markov

models for information retrieval. In *Proceedings of the seventh Text Retrieval Conference, TREC-7*, pages 133–142. NIST Special Publication 500-242, 1999.

[59] D. Miller, T. Leek, and R. Schwartz. A hidden Markov model information retrieval system. In *Proceedings of the 22nd ACM Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 214–221, 1999.

[60] K. Ng. A maximum likelihood ratio information retrieval model. In *Proceedings of the eighth Text Retrieval Conference, TREC-8*. NIST Special Publications, 2000.

[61] NTCIR. NII-NACSIS Test Collection for IR systems, 2005. `http://research.nii.ac.jp/ntcir/index-en.html`.

[62] P. Ogilvie and J. Callan. Combining document representations for known-item search. In C. Clarke, G. Cormack, J. Callan, D. Hawking, and A. Smeaton, editors, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 143–150. ACM Press, New York NY, USA, 2003.

[63] P. Ogilvie and J. Callan. Language models and structured document retrieval. In *Proceedings of the first workshop on the evaluation of XML retrieval (INEX)*, pages 33–40, 2003.

[64] J. Ponte and W. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st ACM Conference on Research and Development in Information Retrieval (SIGIR'98)*, pages 275–281, 1998.

[65] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In A. Waibel and K. Lee, editors, *Readings in speech recognition*, pages 267–296. Morgan Kaufmann, 1990.

[66] H. Rode and D. Hiemstra. Conceptual language models for context-aware text retrieval. In *Proceedings of the 13th Text Retrieval Conference (TREC)*, 2004.

[67] A. Salminen and F. Tompa. PAT expressions: An algebra for text search. In *Proceedings of the 2nd International Conference in Computatio nal Lexicography, COMPLEX'92*, pages 309–332, 1992.

[68] A. Sankar, V. Gadde, A. Stolcke, and F. Weng. Improved modeling and efficiency for automatic transcription of broadcast news. *Speech Communication*, 37:133–158, 2002.

[69] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26:321–343, 1975.

[70] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47, 2002.

[71] C. Shapiro and H. R. Varian. *Information Rules: a strategic guide to the network economy*. Harvard Business School Press, Boston MA, 1999.

[72] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. Blueprint of a cross-lingual web retrieval collection. In R. van Zwol, editor, *Proceedings of the Fifth Dutch-Belgian Information Retrieval Workshop (DIR'5)*, pages 33–38, 2005.

[73] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. Blueprint of a cross-lingual Web retrieval collection. *Journal on Digital Information Management*, 3:9–13, 2005.

[74] F. Song and W. Croft. A general language model for information retrieval. In *Proceedings of the 8th International Conference on Information and Knowledge Management, CIKM'99*, pages 316–321, 1999.

[75] K. Sparck-Jones, S. Robertson, D. Hiemstra, and H. Zaragoza. Language mod-

elling and relevance. In W. Croft and J. Lafferty, editors, *Language Modeling for Information Retrieval*, pages 57–71. Kluwer Academic Publishers, 2003.

[76] A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52:226–234, 2001.

[77] A. Stolcke. Entropy-based pruning of back-off language models. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pages 270–274, 1998.

[78] O. Tsur, M. de Rijke, and K. Sima'an. Biographer: Biography questions as a restricted domain question answering task. In *Proceedings ACL 2004 Workshop on Question Answering in Restricted Domains*, 2004.

[79] C. van Rijsbergen. *Information Retrieval, second edition*. Butterworths, 1979. http://www.dcs.gla.ac.uk/Keith/Preface.html.

[80] E. M. Voorhees. Overview of the TREC 2001 question answering track. In E. M. Voorhees and D. K. Harman, editors, *The Tenth Text REtrieval Conference (TREC 2001)*, pages 42–51. National Institute for Standards and Technology. NIST Special Publication 500-250, 2002.

[81] E. M. Voorhees and L. P. Buckland, editors. *The Eleventh Text REtrieval Conference (TREC 2002)*, 2003. National Institute for Standards and Technology. NIST Special Publication 500-251.

[82] T. Westerveld and A. de Vries. Multimedia retrieval using multiple examples. In *International Conference on Image and Video Retrieval (CIVR'04)*, 2004.

[83] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: compressing and indexing documents and images*. The Morgan Kaufmann series in multimedia information and systems. Morgan Kaufmann Publishers, San Francisco CA, 1999.

[84] J. Xu, R. Weischedel, and C. Nguyen. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th ACM Conference on Research and Development in Information Retrieval (SIGIR'01)*, pages 105–110, 2001.

[85] G.-R. Xue, Q. Yang, H.-J. Zeng, Y. Yu, and Z. Chen. Exploiting the hierarchical structure for link analysis. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05)*, pages 186–193, New York, NY, USA, 2005. ACM Press.

[86] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the 10th International Conference on Information and Knowledge Managemen (CIKM'01)*, pages 403–410, 2001.

[87] Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th ACM Conference on Research and Development in Information Retrieval(SIGIR'02)*, pages 81–88, 2002.