Multilayer Collection Selection and Search of Topically Organized Patents

Michail Salampasis
Inst. of Software Technology
and Interactive Systems
Vienna Univ. of Technology
salampasis@ifs.tuwien.ac.at

Anastasia Giachanou
Dept. of Applied Informatics
University of Macedonia
Thessaloniki, Greece
agiahanou@uom.gr

Georgios Paltoglou
School of Technology
Univ. of Wolverhampton
United Kingdom
g.paltoglou@wlv.ac.uk

Abstract

We present a patent search system that explores three issues: (a) topical organization of patents based on their IPC, (b) collection selection of topically organised patent collections and (c) integration of collection selection tools to patent search systems. Patent documents produced worldwide have manually-assigned classification codes which in our work are used to cluster, distribute and index patents through hundreds or thousands of sub-collections. We propose a new collection selection method suitable for search systems having documents organized using hierarchical classification schemes such as IPC/CPC. The new method uses multiple evidence utilising, for each collection, the ranking of ancestors collections in higher level of the classification hierarchy. We tested our method on CLEF-IP 2011 and compared its performance to state-of-the-art collection selection algorithms. We also integrated this method as a component suggesting patent collections in the iPerFedPat patent search system.

1 Introduction

Distributed Information Retrieval (DIR), also known as federated search [SC03a], offers users the capability of simultaneously searching multiple online remote information sources through a single point of search. The DIR process can be perceived as three separate but interleaved sub-processes: Source representation, in which surrogates of the available remote collections are created [CC01]. Source selection, in which a subset of the available information collections is chosen to process the query [CLC95, SC03b] and results merging, in which the separate results are combined into a single merged result list which is returned to the user [SC03a, PSS08].

Although DIR has been explored for more than 15 years now, it hasn't been widely used in many search applications, mainly because the centralized approach has proved to be scalable and more effective in web search. However, in professional search, DIR could be better applicable and more suitable because quite often high value documents exist that can be naturally organized into sub-collections. Patent search is a very good example because patents have manually-assigned classification codes providing an environment where DIR techniques can be effectively applied. In our study, the International Patent Classification (IPC) codes are used to topically

In: M. Lupu, M. Salampasis, N. Fuhr, A. Hanbury, B. Larsen, H. Strindberg (eds.): Proceedings of the Integrating IR technologies for Professional Search Workshop, Moscow, Russia, 24-March-2013, published at http://ceur-ws.org

cluster, distribute and index patents through hundreds or thousands of sub-collections. Our system automatically selects the best collections/IPCs for each query submitted to the system, something which very precisely and naturally resembles the way patents professionals do various types of patents searches, especially invalidity search.

The work which is presented in this paper is not a typical federated search study, since we focused on logically clustering the patents rather than distributing them at a physical level. We created clusters of patents based on their manually assigned IPC codes and we tested different collection selection methods for the IPC/cluster retrieval task. In that sense our work can be equally seen as a method for cluster-based document retrieval using DIR selection methods. In that context there is no attention given to the claim that DIR can improve the efficiency of patent search. In fact, in case of patent search, where complete patent collections can be acquired easily and the patent collections can be indexed centrally, probably this is not the case at all. On the other hand it should be equally said that because our method is based on clustering the patent documents and not physically distributed them, the DIR methods which we applied (for source selection and results merging) also operate in the scale of milliseconds.

We present a new collection selection method that follows a multilayer, multi-evidence process to suggest collections taking advantage of the special hierarchical classification of patent documents. The proposed method is compared to state-of the-art collection selection algorithms (CORI, BordaFuse, Reciprocal Rank).

Additionally, we explore the issues of integrating an IPC collection selection component to a prototype patent search system called iPerFedPat (www.perfedpat.eu). This system is based on the ezDL [Fuh11] which is a framework for interactive search applications integrating multiple information sources. The iPerFedPat system, based on the ezDL, has a pluggable architecture, providing core services and operations being able to integrate multiple patent data sources and patent related data streams, thus providing multiple patent search tools and UIs. The main utility of integrating the selection component in iPerFedPat is for assisting the retrieval of relevant IPCs during a long-session patent search. In many types of patents search (e.g. prior art search) the core relevant IPCs have already been manually identified during pre-classification before starting the search. However, there are several situations (e.g. when the person conducting a patent search would like to validate the IPCs that have been already assigned or a re-classification is needed), where tools for assisting this task will be very useful.

The rest of this paper is organized as follows. In Section 2 we present in detail how patents are topically organized in our work using their IPC code. In Section 3 we describe the DIR techniques that were tested on patent documents for our study and the new methodology for collection selection proposed in this paper. In Section 4 we describe the details of our experimental setup and the results. We follow with a discussion of the rationale of our approach in Section 5 and future work and conclusions in Section 6.

2 Topically Organised Patents for DIR

In this paper, we extend our previous work of applying DIR methods to topically organized patents [SPG12]. We propose a new collection selection method that surpasses previous source/IPC selection methods for topically organised patents. Another collection selection study involving topically organized patents is reported in the literature [LCC00], however this study was conducted many years ago with a different (USPTO) patent dataset. Also, our approach of dividing patents is different and closer to the actual way of patent examiners conducting patent searches, as we divide patents into a much larger number of sub-collections. Additionally, our approach to apply CORI in multiple layers is new and much more effective.

All patents have manually assigned IPC codes [CC11]. IPC is an internationally accepted standard taxonomy for classifying, sorting, organizing, disseminating, and searching patents. It is officially administered by World Intellectual Property Organization (WIPO). The IPC provides a hierarchical system of language independent symbols for the classification of patents according to the different areas of technology to which they pertain. IPC has currently about 71,000 nodes which are organized into a five-level hierarchical system which is also extended in greater levels of granularity. IPC codes are assigned to patent documents manually by technical specialists.

Patents are classified by a number of different classification schemes. European Classification (ECLA) and U.S. Patent Classification System (USPTO) are the most known classification schemes used by EPO and USPTO respectively. Recently, EPO and USPTO signed a joint agreement to develop a common classification scheme known as Cooperative Patent Classification (CPC). The CPC that has been developed as an extension of the IPC contains over 260,000 individual codes. For this study, patents were organized based on IPC codes because from the available classification schemes in CLEF-IP, IPC is the most widely used and also used by EPO.

Although IPC codes are used to topically cluster patents into sub-collections, something which is a prominent prerequisite for DIR, there are some important differences which motivated us to re-examine and adapt existing

DIR techniques in patent search. Firstly, IPC are assigned by humans in a very detailed and purposeful assignment process, something which is very different by the creation of sub-collections using automated clustering algorithms or the naive division method by chronological or source order, a division method which has been extensively used in past DIR research. Also, patents are published electronically using a strict technical form and structure [Ada10]. This characteristic is another reason to reassess existing DIR techniques because these have been mainly developed for structureless and short documents such as newspapers or poorly structured web documents. Another important difference is that patent search is recall oriented because very high recall is required in most searchers [LMTT11], i.e. a single missed patent in a patentability search can invalidate a newly granted patent. This contrasts with web search where high precision of initially returned results is the requirement and about which DIR algorithms were mostly concentrated and evaluated [PSS08].

Before we describe our study further we should explain IPC which determines how we created the sub-collections in our experiments. Top-level IPC nodes consist of eight sections such as human necessities, performing operations, chemistry, textiles, fixed constructions, mechanical engineering, physics, and electricity. A section is divided into classes which are subdivided into subclasses. Subclass is divided into main groups which are further subdivided into subgroups. In total, the current IPC has 8 sections, 129 classes, 632 subclasses, 7.530 main groups and approximately 63,800 subgroups.

Table 1 shows a part of IPC. Section symbols use uppercase letters A through H. A class symbol consists of a section symbol followed by two-digit numbers like F01, F02 etc. A subclass symbol is a class symbol followed by an uppercase letter like F01B. A main group symbol consists of a subclass symbol followed by one to three-digit numbers followed by a slash followed by 00 such as F01B7/00. A subgroup symbol replaces the last 00 in a main group symbol with two-digit numbers except for 00 such as F01B7/02. Each IPC node is attached with a noun phrase description which specifies some technical fields relevant to that IPC code. Note that a subgroup may have more refined subgroups (i.e. defining 6th, 7th level etc). Hierarchies among subgroups are indicated not by subgroup symbols but by the number of dot symbols preceding the node descriptions as shown in Table 1.

Table 1: An Example of a Section From the IPC Clasification

Section	Mechanical engineering...	F
Class	Machines or engines in general	F01
Subclass	Machines or engines with two or more pistons	F01B
Main Group	reciprocating within same cylinder or...	F01B/7/00
Subgroup	..with oppositely reciprocating pistons	F01B/7/02
Subgroup	..acting on same main shaft	F01B/7/04

3 Collection Selection

3.1 Prior Work on Collection Selection

There are a number of source selection approaches including CORI [CLC95], gGlOSS [FPC$^+$99], and others, that characterize different collections using collection statistics like term frequencies [SJCO02]. These statistics, which are used to select or rank the available collections relevance to a query, are usually assumed to be available from cooperative search providers. Alternatively, statistics can be approximated by sampling uncooperative providers with a set of queries [CC01]. The main characteristic of CORI which is probably the most widely used and tested source selection method is that it creates a hyper-document representing all the documents-members of a sub-collection.

The Decision-Theoretic framework (DTF) presented by Fuhr [Fuh99] is one of the first attempts to approach the problem of source selection from a theoretical point of view. The Decision-Theoretic framework (DTF) produces a ranking of collections with the goal of minimizing the occurring costs, under the assumption that retrieving irrelevant documents is more expensive than retrieving relevant ones.

In more recent years, there has been a shift of focus in research on source selection, from estimating the relevancy of each remote collection to explicitly estimating the number of relevant documents in each. ReDDE [SC03b] focuses at exactly that task. It is based on utilizing a centralized sample index, comprised of all the documents that are sampled in the query-sampling phase and ranks the collections based on the number of documents that appear in the top ranks of the centralized sample index. Its performance is similar to CORI at testbeds with collections of similar size and better when the sizes vary significantly. Two similar approaches named CRCS(l) and CRCS(e) were presented by Shokouhi [Sho07], assigning different weights to the returned documents depending on their rank, in a linear or exponential fashion. Other methods see source selection as a voting method where the available collections are candidates and the documents retrieved from the set of sampled documents are voters [PSS09]. Different voting mechanism can be used (e.g. BordaFuse, Reciprocal Rank, Compsum) mainly inspired by data fusion techniques. The methods described in this paragraph in past DIR experiments attained improvements in precision over previous approaches, but their recall was usually lower.

3.2 Multilayer Collection Selection

We exploit the IPC hierarchical classification scheme and topically organized patents to propose a new multiple-evidence multilayer collection selection method. The new method ranks collections/IPCs not only based on the subdivision of patents in a specific IPC layer, but additionally utilizing the ranking of their ancestors, if the same selection process (query) would had been applied at a higher level. This method can effectively suggest relevant collections at any professional search system where high value documents exist that can be organized hierarchically according to an appropriate classification scheme.

The motivation behind the multilayer method is to select as many as possible relevant collections at low levels (level 4, level 5 etc). IPC code selection when applied at low levels can effectively help patent examiners to identify quickly the subgroups they should focus and this can become a real time saver. In a recent field survey, patent examiners expressed the problem of spending time exploring IPC codes (sub-groups) that discover later they are not relevant. That happens more often in smaller patent offices where patent examiners are usually asked to examine patents in areas which they are knowledgeable but not top experts. In such conditions collection/IPC selection methods and tools could be very useful for patents examiners while searching relevant patents.

The proposed method is based on collections selected by CORI. Previous studies showed that CORI performs better than other collection selection methods (BordaFuse, Reciprocal Rank) when applied at the patent domain [SPG12]. The reason is that CORI is based on a content-based representation of sub-collections using a hyperdocument approach, while the other methods use individual retrieved documents from a sub-collection to estimate the relevance of a sub-collection. However, CORI tends to produce poorer results at low levels (level 4). One reason is that the technological area of patents belonging to a sub-collection is more accurately represented in higher IPC levels as it consists of less sub-collections. At higher IPC levels, documents in one sub-collection are relatively homogeneous and better distinguished from patents in other IPCs, something that is more difficult to capture in lower levels. For example, sub-collections of level 4 that contains about ten times more sub-collections than level 3, are less easier differentiated between each other using a hyperdocument approach, resulting in a decreased CORI performance. To depict this differentiation more clearly, patents that represent methods for dental hygiene can be more easily differentiated from radiation therapy patents (level 3) while patents represent dental machines for boring may not be so easily differentiated from those represent dental tools (level 4).

In order for the algorithm to function using multiple evidence, the documents should be organized in at least two different levels. In this paper, we focus on level 3 (subclass) and level 4 (main group). When a query is submitted to the system, two lists of collections with their relevance scores are returned, one list from level 3 and one from level 4. We used CORI collection selection algorithm to retrieve the relevant collections as it has been proven more effective than other collection selection algorithms (e.g. BordaFuse, Reciprocal Rank) [SPG12].

The lists returned from $level_i$ and $level_{i+1}$ can be represented by two plots using the collection and the score:

$\{(Coll_A, score_A), (Coll_B, score_B), (Coll_C, score_C), ..., (Coll_N, score_N)\}$

$\{(Coll_{A.1}, score_{A.1}), (Coll_{A.2}, score_{A.2}), (Coll_{A.3}, score_{A.3}), ..., (Coll_{A.M}, score_{A.M}), (Coll_{B.1}, score_{B.1}), ..., (Coll_{N.1}, score_{N.1}), ..., (Coll_{N.M}, score_{N.M})\}$

where N is the number of collections suggested at $level_i$, M is the number of collections at $level_{i+1}$ that are children of $collection_A$ and M is the number of collections at $level_{i+1}$ that are children of $collection_N$.

The new collection selection algorithm combines the information gathered from the two levels to produce a new list of relevant collection. The new algorithm evaluates the new scores for collections at $level_{i+1}$ according to the following equation:

$$score_{y.z} = a * score_y + (1 - a) * score_{y.z} \tag{1}$$

where y is a collection at $level_i$ and z is a collection at $level_{i+1}$ which is child of the $Coll_Y$.

The value of parameter a represents the weight of the collections selected at level 3. For our experiments, the value of the weight was decided after a training process. During the training process that preceded the actual runs, we tested various parameters to examine which value optimizes the performance of the method.

4 Experiment

4.1 Experimental Set up

The data collection which was used in the study is CLEF-IP 2011 where patents are extracts of the MAREC dataset, containing over 2.6 million patent documents pertaining to 1.3 million patents from the EPO with content in English, German and French, and extended by documents from the WIPO. We indexed the collection with the Lemur toolkit. The fields which have been indexed are: title, abstract, description (first 500 words), claims, inventor, applicant and IPC class information. Patent documents have been pre-processed to produce a single (virtual) document representing a patent. Our pre-processing involves also stop-word removal and stemming using the Porter stemmer. In our study, we use the Inquery algorithm implementation of Lemur.

We have divided the CLEF-IP collection using the subclass (split3), the main group (split4) and the subgroup level (split5). This decision is driven by the way that patent examiners work when doing patent searches who basically try to incrementally focus into a narrower sub-collection of documents. In the present system, we allocate a patent to each sub-collection specified by at least one of its IPC codes, i.e. a sub-collection might overlap with others in terms of the patents it contains. This is the reason why the column #patents presents a number larger than the 1.3 million patents that constitute the CLEF-IP 2011 collection.

Table 2: Statistics of the CLEF-IP 2011 divisions using different levels of IPC

Split	#patents	Collections Number	Docs per Collection			
			Avg	Min	Max	Median
split3	3622570	632	5732	1	165434	1930
split4	5363045	7530	712	1	83646	144
split5	10393924	63806	163	1	39108	36

To test our system, we used a subset of the official queries provided in CLEF-IP 2011 dataset. We run 50 random queries generated using the title, the abstract, the description and the claims. We tested different combinations of source selection (CORI, BordaFuse, and Reciprocal Rank) at split3 and split4. For results merging, we applied CORI results merging algorithm [CLC95] that is based on a heuristic weighted scores merging algorithm. We also performed runs with the centralized index and the optimal approach for each split. For the optimal run, the system retrieved documents only from the collections containing the relevant documents.

The multilayer method was tested at split4. To test the multilayer method, we used the collections selected by CORI at split3 and split4. Additionally, we run 100 training queries generated randomly to decide which value of the parameter a (equation 1) optimizes the performance of the method. For the experiments in this study, parameter a was assigned the value of 0.8 which means that 80% of the evidence stems from split3.

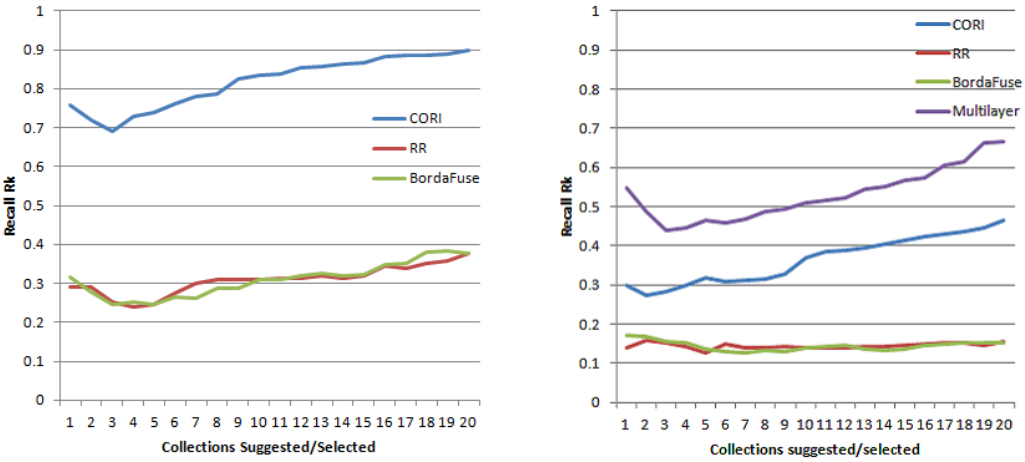
4.2 Results

Source selection algorithms for source recommendation applications (high-recall goal) are compared using a recall measure R_n [CLC95, NF03, Lar03] where the collection ranking provided by the source selection algorithm under investigation is compared with the optimal ranking, under which collections are ranked by the number of relevant documents they possess.

Figure 1 shows the results produced from the source selection algorithms as they gradually select more sub-collections (X axis) at split3 (Figure 1a) and split4 (Figure 1b). The results produced from the multilayer method at split4 are showed on Figure 1b as this method was only tested on split4.

The best performing algorithm at split3 (Figure 1a) is CORI which identifies about 90% of relevant documents in the first 20 suggested collections while the other methods identify less than 40%. At split4 (Figure 1b) the best performing algorithm is the multilayer method where the first 20 suggested collections contain about 67% of all relevant documents while CORI managed to identify 46% of relevant documents. This is a very encouraging result that strongly suggests that source selection algorithms can be effectively used to suggest sub-collections as starting points for information seekers to search. The precision oriented methods Reciprocal Rank and BordaFuse produce poor results since they manage to identify only 15% of relevant documents in the first 20 selected sub-collections.

Table 3 shows the results from the runs performed on the centralized index and the top run for each DIR method that was applied on the split3 and split4.The multilayer was performed and evaluated at split4. The



(a) Level of subclass

(b) Level of main group

Figure 1: Results of the source suggestion algorithm(s)

performance of the methods can be also compared to the performance of the optimal run.

Table 3: Results of run of each DIR method at split3

Run Description	split3			spit4		
	PRES @100	MAP @100	RECALL @100	PRES @100	MAP @100	RECALL @100
Optimal	0,311	0,116	0,397	0,33	0,117	0,399
Centralized	0,264	0,113	0,345	0,264	0,113	0,345
10-100.CORI-CORI	0,291	0,103	0,363	0,19	0,052	0,262
10-100.BordaFuse-CORI	0,102	0,039	0,143	0,076	0,022	0,115
10-100.RR-CORI	0,1	0,039	0,128	0,073	0,023	0,102
10-100.Multilayer	-	-	-	0,269	0,086	0,32

5 Discussion

The results show that the best performing source selection algorithm at the level of subclass (split3) is CORI. The superiority of CORI as source selection method compared to BordaFuse and Reciprocal Rank is unquestionable and consistent with our previous study [SPG12]. However, we observe that as the number of sub-collections increases (i.e. IPC level 4, 5 etc), the performance of the collection selection algorithms deteriorates in contrast to the optimal method that performs better at lower levels (level4).

The most interesting and important finding for this study is that the multilayer method performs better than the other methods at lower levels. The multilayer method managed to select more relevant collections than CORI at split4 by utilising information from previous levels. The performance of actual runs using our source selection method at split4 is better than using CORI as source selection but also when compared to the centralized index approach. Additionally, it is very interesting that some DIR approaches managed to perform better than the centralized approach that is also an assumption from a previous study [SPG12]. This finding shows that DIR approaches not only can be more efficient and probably more appropriate due to the dynamic nature of creating documents in the patent domain, but also more effective.

It seems that DIR methods, at least in patent search, can be applied in a way resembling more the cluster-based approaches to information retrieval [Wil88, FLSG12] and could improve efficiency and effectiveness. As for efficiency, searching and browsing on sub-collections rather than the complete collection could significantly reduce the retrieval time and more significantly the information seeking time of users. In relation to effectiveness, the potential of DIR retrieval stems from the cluster hypothesis [Rij79] which states that related documents residing in the same cluster (sub-collection) tend to satisfy same information needs. The clustering hypothesis was proved by Fuhr [FLSG12] who developed the optimum clustering framework. The expectation in the context of source selection, which is of primarily importance for this study, is that if the correct sub-collections are selected then it will be easier for relevant documents to be retrieved from the smaller set of available documents and more effective searches can be performed.

The field of DIR has been explored in the last decade mostly as a response to technical challenges such as the prohibitive size and exploding rate of growth of the web which make it impossible to be indexed completely [RGM01]. Also there is a large number of online sources (web sites), known as invisible web which are not reachable by search engines. As the main focus of this paper is patent search, we should mention this is especially true in the patent domain as nearly all authoritative online patent sources (e.g. EPOs espacenet) are not indexable and therefore not accessible by general purpose search engines.

From our one-to-one interviews in a small patent office with patent examiners, the majority of them said that an IPC suggestion tool may be useful in their searches. Therefore we integrated the IPC selection method as a tool in the iPerFedPat system (www.perfedpat.eu) for suggesting relevant IPCs. Such tool can be a time saver for patent examiners as they can focus their search to a narrow set of relevant collections. Additionally, of high importance is that patent examiners will have the opportunity to use the tool in a combination with other tools of the system such as searching specific datasets resulting in a faster and more efficient patent search process.

The integration of the IPC suggestion tool was implemented sending http requests to an external server providing the IPC selection services. The server receives the requests and sends a response back about the IPCs suggested. In iPerFedPat there are more search tools integrated in similar way (tools for faceted search, entity extraction, clustering search). From a information seeking process perspective, the integration of different search tools with the main retrieval engine (producing ranked lists of patent documents in response to a query), allows different search interfaces to coexist in an information seekers patent search system. This client-server integration provides the core services to the patents search system but synchronization between the tools is required so one event or action in one tool (for example selecting an IPC) can update the views produced from the main retrieval engine or the other tools. We plan to extend the integration scheme of iPerFedPat with a communication and coordination language to address this need.

6 Conclusion and Future Work

In this paper, a new collection selection algorithm was presented for sub-collections divided using a hierarchical classification scheme. To test the new collection selection method, we divided the CLEF-IP collection into clusters using the subclass (split3) and the main group (split4) level to experiment with different levels and depth of topical organization. The new method was compared with state-of-the-art algorithms and the centralized approach.

The results showed that the best performing source selection algorithm was the multilayer method at the level of main group. On the other side the (precision oriented) methods Reciprocal Rank and BordaFuse consistently produced worse results.

We plan to continue this work. One issue which we wish to explore further is how the collection selection methods would perform if an automatic or a semi-automatic clustering method would be applied. We would also like to explore how features such as sub-collections sizes could influence the performance of the multilayer collection selection method. Also, we would like to experiment with larger distribution levels based on IPC (subgroup level). We produced divisions of higher granularity at level 5 of IPC but we didnt have the time and the resources to report results for this division (split5). We plan to report the runs using split5 in a future paper.

We also explored issues related to the integration of an IPC suggestion tool to a patent search system. However, we know that user-centered studies are needed and are more appropriate to decide the usefulness of such tools and we plan to conduct them in the near future.

In conclusion, we feel that the discussion and the experiment presented in this paper are useful to the designers of patent search systems which are based on DIR methods that were more effective and efficient than others which are based on centralized approaches. Of course, more and larger experiments are required before we can reach a more general conclusion. However, our experiment has produced some indications advocating the development of patent search systems which would be based on similar principles with the ideas that inspired the adaptation and use of DIR methods and their integration in patent search systems.

Acknowledgements

The second author is supported by a Marie Curie fellowship from the IEF project PerFedPat (www.perfedpat.eu).

References

[Ada10] Stephen Adams. The text, the full text and nothing but the text: Part 1 standards for creating textual information in patent documents and general search implications. *World Patent Information*, 32(1):22 – 29, 2010.

[CC01] Jamie Callan and Margaret Connell. Query-based sampling of text databases. *ACM Transactions on Information Systems*, 19(2):97–130, April 2001.

[CC11] Yen-Liang Chen and Yu-Ting Chiu. An ipc-based vector space model for patent retrieval. *Information Processing I& Management*, 47(3):309 – 322, 2011.

[CLC95] James P. Callan, Zhihong Lu, and W. Bruce Croft. Searching distributed collections with inference networks. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '95, pages 21–28, New York, NY, USA, 1995. ACM.

[FLSG12] Norbert Fuhr, Marc Lechtenfeld, Benno Stein, and Tim Gollub. The optimum clustering framework: implementing the cluster hypothesis. *Information Retrieval*, 15(2):93–115, April 2012.

[FPC$^+$99] James C. French, Allison L. Powell, Jamie Callan, Charles L. Viles, Travis Emmitt, Kevin J. Prey, and Yun Mou. Comparing the performance of database selection algorithms. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 238–245, New York, NY, USA, 1999. ACM.

[Fuh99] Norbert Fuhr. A decision-theoretic approach to database selection in networked ir. *ACM Trans. Inf. Syst.*, 17(3):229–249, 1999.

[Fuh11] Norbert Fuhr. An infrastructure for supporting the evaluation of interactive information retrieval. In *Proceedings of the 2011 workshop on Data infrastructurEs for supporting information retrieval evaluation*, DESIRE '11, pages 1–2, New York, NY, USA, 2011. ACM.

[Lar03] Ray R. Larson. Distributed ir for digital libraries. In *In Research and Advanced Technology for Digital Libraries (ECDL 2003*, pages 487–498. Springer (LNCS, 2003.

[LCC00] Leah S. Larkey, Margaret E. Connell, and Jamie Callan. Collection selection and results merging with topically organized u.s. patents and trec data. In *Proceedings of the ninth international conference on Information and knowledge management*, CIKM '00, pages 282–289, New York, NY, USA, 2000. ACM.

[LMTT11] Mihai Lupu, Katja Mayer, John Tait, and Anthony J. Trippe. *Current Challenges in Patent Information Retrieval.* The Information Retrieval Series. Springer, 2011.

[NF03] Henrik Nottelmann and Norbert Fuhr. Evaluating different methods of estimating retrieval quality for resource selection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 290–297, New York, NY, USA, 2003. ACM.

[PSS08] Georgios Paltoglou, Michail Salampasis, and Maria Satratzemi. A results merging algorithm for distributed information retrieval environments that combines regression methodologies with a selective download phase. *Information Processing & Management*, 44(4):1580–1599, July 2008.

[PSS09] Georgios Paltoglou, Michail Salampasis, and Maria Satratzemi. Simple adaptations of data fusion algorithms for source selection. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy, editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 497–508. Springer Berlin Heidelberg, 2009.

[RGM01] Sriram Raghavan and Hector Garcia-Molina. Crawling the hidden web. In *Proceedings of the 27th International Conference on Very Large Data Bases*, VLDB '01, pages 129–138, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[Rij79] C. J. Van Rijsbergen. *Information Retrieval.* Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[SC03a] Luo Si and Jamie Callan. A semisupervised learning method to merge search engine results. *ACM Transactions on Information Systems*, 21(4):457–491, 2003.

[SC03b] Luo Si and Jamie Callan. Relevant document distribution estimation method for resource selection. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03*, page 298, 2003.

[Sho07] Milad Shokouhi. Central-rank-based collection selection in uncooperative distributed information retrieval. In *Proceedings of the 29th European conference on IR research*, ECIR'07, pages 160–172, Berlin, Heidelberg, 2007. Springer-Verlag.

[SJCO02] Luo Si, Rong Jin, Jamie Callan, and Paul Ogilvie. A language modeling framework for resource selection and results merging. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pages 391–397, New York, NY, USA, 2002. ACM.

[SPG12] Michail Salampasis, Georgios Paltoglou, and Anastasia Giahanou. Report on the clef-ip 2012 experiments: Search of topically organized patents. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[Wil88] Peter Willett. Recent trends in hierarchic document clustering: a critical review. *Inf. Process. Manage.*, 24(5):577–597, August 1988.