Report on the CLEF-IP 2013 Experiments: Multilayer Collection Selection on Topically Organized Patents

Anastasia Giachanou[1] Michail Salampasis[2] Maya Satratzemi[1] and Nikolaos Samaras[1]

[1] *University of Macedonia, Department of Applied Informatics, Thessaloniki, Greece*
[2] *Vienna University of Technology, Institute of Software Technology and Interactive Systems, 1040, Vienna, Austria*

agiahanou@uom.gr, salampasis@ifs.tuwien.ac.at, maya@uom.gr, samaras@uom.gr

Abstract. This technical report presents the work which has been carried out using Distributed Information Retrieval methods for federated search of patent documents for the passage retrieval starting from claims (patentability or novelty search) task. 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. For source selection, we tested CORI and a new collection selection method, the Multilayer method. We also tested CORI and SSL results merging algorithms. We run experiments using different combinations of the number of collections requested and documents retrieved from each collection. One of the aims of the experiments was to test older DIR methods that characterize different collections using collection statistics like term frequencies and how they perform in patent search and in suggesting relevant collections. Also to experiment with Multilayer, 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. We submitted 8 runs. According to PRES @100 our best DIR approach ranked 6th across 21 submitted results.

Keywords: Patent Search, IPC, Source Selection, Federated Search

1 Introduction

This technical report presents the participation of the University of Macedonia in collaboration with the Vienna University of Technology in the passage retrieval (patentability or novelty search) task. Our experiments aim to explore an important issue, the thematic organization of patent documents using the subdivision of patent data by International Patent Classification (IPC) codes, and if this organization can be used to improve patent search effectiveness using DIR methods in comparison to centralized index approaches. We have also developed and tested 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.

Patent documents produced worldwide have manually-assigned classification codes which in our experiments are used to topically organize, distribute and index patents through hundreds or thousands of sub-collections. Our system automatically selects the best collections 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 patent examiners doing invalidity search.

In the experiments which are reported in this paper, we divided the CLEF-IP collection using the subclass (Split-3), the main group (Split-4) and the subgroup level (Split-5). The patents have been allocated to sub-collections based on the IPC codes specified in them. In the experiments we report here, we allocated a patent to each sub-collection specified by at least one of its IPC code, i.e. a sub-collection might overlap with others in terms of the patents it contains.

Topics in the patentability or novelty search task are sets of claims extracted from actual patent application documents. Participants are asked to return passages that are relevant to the topic claims. The topics contain also a pointer to the original patent application file. Our participation was limited only at the document level. We didn't perform the claims to passage task because the main objective of our method is to identify relevant IPCs. We submitted 8 runs. According to PRES @100 our best submitted DIR approach ranked 6th across 21 submitted results.

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

The experiments which are reported in this paper extend our previous work of applying DIR methods to topically organized patents (Salampasis et al. 2012). 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 (Larkey et al. 2000), 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, we apply CORI in multiple layers and evaluate its performance.

All patents have manually assigned IPC codes (Chen & Chiu 2011). 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 can be 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 this was the available classification scheme in the test collection CLEF-IP.

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. Also, patents are published electronically using a strict technical form and structure (Adams 2010). 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 (Lupu et al. 2011), 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 (Paltoglou et al. 2008).

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 which are: 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 Classification

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 …	F01B7/00
Subgroup	.with oppositely reciprocating pistons	F01B7/02
Subgroup	..acting on same main shaft	F01B7/04

3 Distributed IR on Patent Search

3.1 Prior Work on Collection Selection

Distributed Information Retrieval (DIR), also known as federated search (Si & J. Callan 2003a), 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 (Callan & Connell 2001). Source selection, in which a subset of the available information collections is chosen to process the query (Paltoglou et al. 2011) and results merging, in which the separate results returned from remote collections are combined into a single merged result list which is returned to the user for examination (Paltoglou et al. 2008).

There are a number of Source Selection approaches including CORI (Callan et al. 1995), gGlOSS (French et al. 1999), and others (Si et al. 2002), that characterize different collections using collection statistics like term frequencies. 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 (Callan & Connell 2001).

The Decision-Theoretic framework (DTF) presented by Fuhr (Fuhr 1999) 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 (Si & Callan 2003b) focuses at exactly that purpose. 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. Other methods see source selection as a voting method where the available collections are candidates and the documents that are retrieved from the set of sampled documents are voters (Paltoglou et al. 2009). Different voting mechanism can be used (e.g. BordaFuse, ReciRank, Compsum) mainly inspired by data fusion techniques.

There is a major difference between CORI and the other collection selection algorithms presented in the paragraph above. CORI builds a hyperdocument representing the sub-collection while using the other methods the collection selection or not is based on the retrieval or not of individual documents from the single centralized sample index. Due to this characteristic CORI may not work well in environments containing a mix of "small" and "very large" document databases. On the other hand for homogenous sub-collections as the ones produced from patents belonging to a single IPC, representative hyperdocument in CORI should normally encompass a strong discriminating power, something useful for effective and robust collection selection.

In the experiments which we report in this paper we use both CORI and our Multilayer method which adapts the way any selection method (CORI in the experiments presented here) can be applied in patent domain.

3.2 Multilayer Collection Selection

We exploit the hierarchical organization of the IPC classification scheme and the idea of 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 utilizes the ranking of their ancestors, if the same selection process (query) 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 are organized hierarchically according to an appropriate classification scheme.

The motivation behind the Multilayer method is to select as many as possible relevant collections at lower IPC 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 we conducted, 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 relatively knowledgeable but not experts. Especially 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 (Salampasis et al. 2012; Giachanou et al. 2012). We believe 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 IPC levels (level 4 or level 5). One reason is that the technological area of patents belonging to a sub-collection is more accurately represented in higher IPC levels (e.g. subclass) because 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

of classification. 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 oral or dental hygiene can be more easily differentiated from radiation therapy patents at level 3 while patents represent dental machines for boring may not be so easily differentiated from patents that represent dental tools at level 4.

In other words our method to apply source selection introduces a normalisation procedure which takes into account the source selection results at several classification levels. Of course, the proposed method can utilise multiple evidence, if the documents are organized in at least two different levels. In this paper, we focus on level 3 (subclass), level 4 (main group) and level 5 (subgroup). We used the CORI collection selection algorithm to retrieve the relevant collections as it has been proven more effective than other collection selection algorithms (e.g. BordaFuse, RR) that we tested before (Salampasis et al. 2012; Giachanou et al. 2012).

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

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

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

where N is the number of sub-collections suggested at $level_i$, M is the number of sub-collections at $level_{i+1}$ that are children of $collection_A$ and M' is the number of sub-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 collections. The new algorithm evaluates the new scores for collections at $level_{i+1}$ according to the following equation:

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

where y is a sub-collection at $level_i$ and z is a sub-collection at $level_{i+1}$ which is child of the $Coll_Y$. Parameter α determines the weight that each level will take to decide the final score of a sub-collection (IPC).

Another parameter of our method is the *collection window* which represents the number of sub-collections that will be re-ranked after taking evidence from a higher level. For example if the aim is to produce a list of N suggested IPCs at level 5, the method should define how many IPCs in the initial rank produced by running CORI in level 5, initially positioned after position N, will be reconsidered in the second round re-ranking process. This is the *window* parameter and this decision can be based either on a fixed threshold such as 100 or on a number relative to the number of IPCs that should be suggested (i.e. 2 * N, 3 * N etc). Another parameter of our method is *influence factor*, i.e. how many IPCs from a higher level (level 4 in our example) should be used to re-rank the *collection window* IPCs in the lower level

(level 5). For example, if we want to re-rank 2*N IPCs in level 5, a parameter in our method is how many IPCs from level 4 we will use to make the re-ranking.

For the experiments in this study, we decided to use the parameters that optimized the performance of Multilayer in a previous study (Giachanou et al. 2012). Based on the results of that study, we decided the parameter *a* to be assigned with the value of 0.8. Finally, at split-4 the collection window and the influence factor were assigned with the values of 20 and 200 respectively while at split-5 those parameters were assigned with the values of 200 and 2000. The decision was based on a previous study that was performed on CLEF 2012.

4 Experimental Setup

The data collection which was used in the study is CLEF-IP 2013 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 sub-group 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
split_3	3622570	632	5732	1	165434	1930
split_4	5363045	7530	712	1	83646	144
split_5	10393924	63806	163	1	39108	36

To test our system, we used a subset of the official queries provided in CLEF-IP 2013 dataset. The queries generated using the title, the abstract, the description and the claims. Topics in French and German were first translated in English using the WIPO Translation Assistant[1]. We tested CORI and Multilayer source selection methods at split3, split4 and split5. For results merging, we applied CORI results merging algorithm (Callan et al. 1995) that is based on a heuristic weighted scores merging

[1] https://www3.wipo.int/patentscope/translate/translate.jsf

algorithm and SSL. We also performed a run with the centralized index. The multi-layer method was tested at split4 and split5. To test the Multilayer method, we used the collections selected by CORI at split3, split4 and split5.

5 Results and Discussion

Table 3 shows the submitted runs ranked according to PRES @100. In each line the experiment description encodes: the number of collections selected, number of documents requested from each selected collection, how patents were topically organized (split-3, split-4 or split-5), method for source selection, method for merging and set of queries (English, all). There are also lines that show the average and median values of the submitted runs of the rest teams. The average and median values were calculated after removing the outliers (the top and the last run according to PRES @100). We should also mention that at the moment writing this report, we are only aware of the results and not of the methods that were used.

Table 3. Results of the submitted runs

Run description	PRES @100	Recall @100	Map @100
centralised.EN	0.420	0.519	0.170
10-100.CORI.SSL.split5.EN	0.418	0.504	0.180
10-100.CORI.CORI.split3.EN	0.414	0.496	0.172
20-50.CORI.CORI.split5.EN	0.414	0.501	0.184
10-100.Multilayer.CORI.split4.EN	0.413	0.515	0.153
20-50.Multilayer.CORI.split5.EN	0.396	0.464	0.178
Median (us NOT including)- EN	0.39	0.488	0.166
Average (us NOT including)- EN	0.385	0.482	0.161
10-100.Multilayer.CORI.split5.EN	0.373	0.451	0.160
10-100.CORI.SSL.split4.EN	0.355	0.431	0.152
Average (us NOT including) - all	0.246	0.31	0.109
10-100.CORI.CORI.split3.all	0.240	0.292	0.099
10-100.Multilayer.CORI.split4.all	0.238	0.313	0.090
centralised.all	0.236	0.293	0.100
20-50.CORI.CORI.split5.all	0.236	0.309	0.101
Median (us NOT including) - all	0.236	0.296	0.114
10-100.CORI.SSL.split5.all	0.230	0.305	0.097
20-50.Multilayer.CORI.split5.all	0.209	0.262	0.094
10-100.Multilayer.CORI.split5.all	0.189	0.245	0.075
10-100.CORI.SSL.split4.all	0.150	0.202	0.057

As it is shown in Table 3 Multilayer performs better than the CORI at the main group (Split4) level. To obtain a more complete picture of the results, we calculated a

recall measure R_n which is used to compare the performance of source selection algorithms (Callan et al. 1995; Nottelmann & Fuhr 2003; Larson 2003).

Table 4 shows the results produced from the source selection algorithms ranked according to R_k @10, R_k @20 and R_k @50 at split4 and split5. The best performing algorithm at split4 is the Multilayer method where the first 50 suggested collections contain about 70% of all relevant documents while CORI managed to identify about 45%. 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.

Table 4. Analysis of IPC distribution of topics and their relevant documents

Source Selection Algorithm -Split	**R_k @10**	**R_k @20**	**R_k @50**
Split 4			
CORI	0.26	0.318	0.459
Multilayer	0.463	0.546	0.693
Split 5			
CORI	0.366	0.369	0.468
Multilayer	0.335	0.346	0.507

Another important finding is that the best runs are those requesting fewer sub-collections (10 collections) and more documents from each selected sub-collection. This fact is probably the result of the small number of relevant documents which exist for each topic. To validate these observations we did a post-run analysis of the topics and how their relevant documents are allocated to sub-collections in each split (Table 5). Table 5 reveals useful information which shows that to some extend relevant IPC codes can be effectively identified if IPC classification codes are already assigned to a topic. This is a feature that we didn't use in our experiments and can be used as a heuristic that could substantially increase the performance of source selection.

Table 5. Analysis of IPC distribution of topics and their relevant documents

IPC Level - Split	# relevant docs per topic (a)	# of IPC classes of each topic (b)	# of IPC classes of relevant docs (c)	c/b	# of common IPC classes between (b) and (c)
Split 3					
ALL	3.35	2.76	4.72	1.71	1.8
EN ONLY	3.87	3.37	5.96	1.77	2.5
Split 4					
ALL	3.35	4.5	7.53	1.67	2.3
EN ONLY	3.87	5.04	10.3	2.04	3.35
Split 5					
ALL	3.35	7.03	17.4	2.48	3.43
EN ONLY	3.87	7.76	21.3	2.74	4.52

In addition to the comments already discussed, perhaps the most interesting and important finding for this study is that DIR approaches managed to perform similar or better than the centralized index approaches. It is also very interesting that the performance remains relatively high at subgroup level (split-5), the level that patent examiners focus on. This is a very interesting finding which shows that DIR approaches can be used to suggest collections at low levels while being effective and efficient.

It seems that in patent domain the cluster-based approaches to information retrieval (Willett 1988)(Fuhr et al. 2012) which utilize document clusters (sub-collections), could be utilized so efficiency or effectiveness can be improved. As for efficiency, searching and browsing on sub-collections rather than the complete collection of documents could significantly reduce the retrieval time of the system and more significantly the information seeking time of users. In relation to effectiveness, the potential of DIR retrieval stems from the cluster hypothesis (Van Rijsbergen 1979) which states that related documents residing in the same cluster (sub-collection) tend to satisfy same information needs. The cluster hypothesis has been utilized in various settings for information retrieval such as for example cluster-based retrieval, extensions of IR models with clusters, latent semantic indexing. 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 focused 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 (Raghavan & Garcia-Molina 2001). Also there is a large number of online sources (web sites), collectively known as invisible web which are either not reachable by search engines because they sit behind pay-to-use turnstiles, or for other reasons do not allow their content to be indexed by web crawlers, offering their own search capabilities (Miller 2007). 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. EPO's espacenet) are not indexable and therefore not accessible by general purpose search engines.

6 Conclusion and Future Work

In this paper we presented the work which has been carried out using Distributed Information Retrieval methods for federated search of patent documents for the CLEF-IP 2013 passage retrieval starting from claims (patentability or novelty search) task. We tested CORI and Multilayer methods for source selection and CORI and SSL for results merging. We have divided the CLEF-IP collection using the subclass (Split-3), the main group (Split-4) and the subgroup (Split-5) level to experiment with different levels and depth of topical organization.

We submitted 8 runs. According to PRES @100 our best DIR approach ranked 6th across 21 submitted results. The methods we apply performed similar or better than the centralised approach.

We plan to explore further this line of work with exploring modifications to the Multilayer and to make it more effective for patent search. We believe that the discussion and the experiment presented in this paper are also useful to the designers of patent search systems which are based on DIR methods.

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

7 References

1. Adams, S., 2010. 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), pp.22–29.
2. Callan, J P, Lu, Z & Croft, W B, 1995. Searching distributed collections with inference networks. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 21–28.
3. Callan, J. & Connell, M., 2001. Query-based sampling of text databases. *ACM Transactions on Information Systems*, 19(2), pp.97–130.
4. Callan, James P, Lu, Zhihong & Croft, W Bruce, 1995. 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*. Seattle, Washington: ACM New York, NY, USA, pp. 21–28.
5. Chen, Y.-L. & Chiu, Y.-T., 2011. An IPC-based vector space model for patent retrieval. *Information Processing & Management*, 47(3), pp.309–322.
6. French, J.C. et al., 1999. Comparing the Performance of Database Selection Algorithms. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 99*, pp.238–245.
7. Fuhr, N., 1999. A decision-theoretic approach to database selection in networked IR. *ACM Transactions on Information Systems*, 17(3), pp.229–249.
8. Fuhr, N. et al., 2012. The optimum clustering framework: implementing the cluster hypothesis. *Information Retrieval*, 15(2), pp.93–115.
9. Giachanou, A., Salampasis, M & Paltoglou, G, 2012. Multilayer Collection Selection and Search of Topically Organized Patents. In *ceur-ws.org*.
10. Larkey, L.S., Connell, M.E. & Callan, J., 2000. 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*. McLean, Virginia, USA: ACM New York, NY, USA, pp. 282–289.
11. Larson, R., 2003. Distributed IR for digital libraries T. Koch & I. Sølvberg, eds. ... *and Advanced Technology for Digital Libraries*, 2769, pp.487–498.
12. Lupu, M. et al., 2011. *Current Challenges in Patent Information Retrieval* M. Lupu et al., eds., Springer.
13. Miller, J., 2007. Most fed data is un-googleable. *Federal ComputerWeek*.
14. Nottelmann, H. & Fuhr, N., 2003. 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. Toronto, Canada: ACM New York, NY, USA, pp. 290–297.
15. Paltoglou, G, Salampasis, M & Satratzemi, M, 2008. A results merging algorithm for distributed information retrieval environments that combines regression methodologies with a selective download phase. *Information Processing & Management*, 44(4), pp.1580–1599.
16. Paltoglou, G., Salampasis, M. & Satratzemi, M., 2008. A results merging algorithm for distributed information retrieval environments that combines regression methodologies with a selective download phase. *Information Processing & Management*, 44(4), pp.1580–1599.
17. Paltoglou, G., Salampasis, M. & Satratzemi, M., 2009. *Advances in Information Retrieval* M. Boughanem et al., eds., Berlin, Heidelberg: Springer Berlin Heidelberg.
18. Paltoglou, G., Salampasis., M & Satratzemi, M., 2011. Modeling information sources as integrals for effective and efficient source selection. *Information Processing & Management*, 47(1), pp.18–36.
19. Raghavan, S. & Garcia-Molina, H., 2001. Crawling the Hidden Web. In *Proceedings of the International Conference on Very Large Data Bases*. Citeseer, pp. 129–138.
20. Van Rijsbergen, C.J., 1979. *Information Retrieval*, Butterworths.
21. Salampasis, M., Paltoglou, G. & Giahanou, A., 2012. Report on the CLEF-IP 2012 Experiments: Search of Topically Organized Patents. In P. Forner, J. Karlgren, & C. Womser-Hacker, eds. *CLEF (Online Working Notes/Labs/Workshop)*.
22. Si, L et al., 2002. A language modeling framework for resource selection and results merging. In *ACM CIKM 02*. ACM Press, pp. 391–397.
23. Si, Luo & Callan, J., 2003a. A semisupervised learning method to merge search engine results. *ACM Transactions on Information Systems*, 21(4), pp.457–491.
24. Si, Luo & Callan, J., 2003b. Relevant document distribution estimation method for resource selection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03*. Toronto, Canada: ACM New York, NY, USA, pp. 298–305.
25. Willett, P., 1988. Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management*, 24(5), pp.577–597.