Report on the CLEF-IP 2012 Experiments: Search of Topically Organized Patents

Michail Salampasis

Vienna University of Technology
Institute of Software Technology and Interactive Systems
Vienna, Austria
salampasis@ifs.tuwien.ac.at

Giorgos Paltoglou

University of Wolverhampton
School of Technology
United Kingdom
g.paltoglou@wlv.ac.uk

Anastasia Giahanou

University of Macedonia
Department of Applied Informatics
Thessaloniki, Greece
agiahanou@gmail.com

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. We tested different combinations of source selection (CORI, BordaFuse, Reciprocal Rank) and results merging algorithms (SSL, CORI). We also tested 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. Also to experiment with newer DIR methods which focus on explicitly estimating the number of relevant documents in each collection and usually attain improvements in precision over previous approaches, but their recall is usually lower. However, the most important aim was to examine how DIR methods will perform if patents are topically organized using their IPC and if DIR methods can approximate the performance of a centralized index approach. We submitted 8 runs. According to PRES @100 our best DIR approach ranked 7th across 31 submitted results, however our best DIR (not submitted) run outperforms all submitted runs.

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

1 Introduction

This technical report presents the participation of the Vienna University of Technology in collaboration with the University of Wolverhampton and the University of Macedonia in the passage retrieval starting from claims (patentability or novelty search) task. Our experiments which are reported here 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.

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. IPC is a standard taxonomy for classifying patents, and 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 is consistently maintained by an authorized official organization, and IPC codes are assigned to patent documents manually by technical specialists such as patent examiners.

In the experiments which are reported in this paper, we divided the CLEF-IP collection using the Subclass (Split-3) and the main group (Split-4) level. The patents have been allocated to sub-collections based on the IPC codes specified in them. The average patent has three IPC codes. 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 and we didn’t perform the claims to passage task. We run two different sets of queries. One set was based on the specific claim(s) which is listed for each topic and the other on the full set of claims which could be extracted from the original patent application file. We submitted 8 runs but in this technical report we report on all the runs we performed and the results we obtained. According to PRES @100 our best submitted DIR approach ranked 7th across 31 submitted results, however our best DIR (not submitted) run outperforms all submitted runs.

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 we applied and the details of our experimental setup. In Section 4 we report the results of our experiments. 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

Distributed Information Retrieval (DIR), also known as federated search (Luo Si & 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 (J. Callan & Connell, 2001). Source selection, in which a subset of the available information collections is chosen to process the query (Georgios Paltoglou, Salampasis, & Satratzemi, 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 (G Paltoglou, Salampasis, & Satratzemi, 2008).

The experiments which are reported in this paper measure the effectiveness of DIR methods when applied to topically organized patents. Another collection selection study involving topically organized patents is reported in the literature (Larkey, Connell, & Callan, 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. We also use state-of-the art algorithms which were not available in the study reported by Larkey.

All patents have manually assigned international patent classification (IPC) codes (Chen & Chiu, 2011). IPC is an internationally accepted standard taxonomy for sorting, organizing, disseminating, and searching patents. It is officially administered by World Intellectual Property Organization. The International Patent Classification (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 is a standard taxonomy for classifying patents, and 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 is consistently maintained by an authorized official organization, and IPC codes are assigned to patent documents manually by technical specialists such as patent examiners.

The fact that 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, raise the necessity to reassess existing DIR techniques in the patent domain. 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 and the last years has been given to web documents. Another important difference is that patent search is recall oriented because very high recall is required in most searchers (Lupu, Mayer, Tait, & Trippe, 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 (G 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 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, and 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 such as F01, F02 etc. A subclass symbol is composed of 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 and B64C781/00. 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	F01B7
Main group	reciprocating within same cylinder or …	F01B7/00
Subgroup	.with oppositely reciprocating pistons	F01B7/02
Subgroup	..acting on same main shaft	F01B7/04

The taxonomy and set of classes, subclasses, groups etc is dynamic. The patent office tries to keep membership to groups down to a maximum by making new subgroups etc. However, new patent applications/inventions require the continual update of the IPC taxonomy. Periodically, the WIPO carries out a reclassification. Sometimes existing subclasses/groups/subgroups are subdivided into new subsets. Sometimes a set of subclasses of a class are merged together, then subdivided again in a different manner. After new subclasses are formed, the patents involved may or may not be assigned to the new subclasses.

3 Experiment

3.1 Setup

The data collection which we have used in the study presented in this paper is CLEF-IP 2012 where patents are extracts of the MAREC dataset, containing over 2.6 million patent documents pertaining to 1.3 million patents from the European Patent Office 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 the experiments reported here we use the Inquery algorithm implementation of Lemur. Although we have submitted results for the French and German topics for reason of completeness, we focused our interest to the English subset of the topics. Also, we didn't perform the passage retrieval task but we focused only in the document retrieval i.e. given a topic/set of claims we tried to retrieve relevant documents in the collection.

We have divided the CLEF-IP collection using the Subclass (Split-3), the main group (Split-4) level and the sub-group level (Split-5). 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. The patents have been allocated to sub-collections based on the IPC codes specified in them. The average patent has three IPC codes. In the present system, we allocate 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, This is the reason why the column #patents presents a number larger than the 1.3 million patents that constitute the CLEF-IP 2012 collection. We could also attempt only to place patents into their unique original reference subgroup (the main IPC code) but this information is not always available in all patent documents.

Table 2. Statistics of the CLEF-IP 2012 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

3.2 DIR methods used

There are a number of Source Selection approaches including CORI (J. P. Callan, Lu, & Croft, 1995), gGlOSS (French et al., 1999), and others (L Si, Jin, Callan, & Ogilvie, 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 (J. Callan & Connell, 2001). 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.

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 (Luo 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. Two similar approaches named CRCS(l) and CRCS(e) were presented by (Shokouhi, 2007), 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 that are retrieved from the set of sampled documents are voters (Georgios Paltoglou, Salampasis, & Satratzemi, 2009). Different voting mechanism can be used (e.g. BordaFuse, ReciRank, 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.

We did an adaptation to source selection algorithms which are used in the experiments. Usually the precision oriented source selection methods such as ReciRank use few documents (e.g. 50) from the sampling collection to estimate the relevancy of remote collections. Based on previous test runs which produced poor results we choose to use 1000 documents from the sampling collection to decide on resources (IPC-based sub-collections) relevancy.

In the experiments which are reported in this paper we used both CORI and the precision oriented methods for source selection with the voting method using Reciprocal Rank and BordaFuse (Aslam & Montague, 2001)

We run the experiments with two different set of queries, one set was based on the specific claim(s) that was listed for each topic (spClms) and the other on the full set of claims which could be extracted from the original patent application file (AllClms). We tested different combinations of source selection (CORI, BordaFuse, and Reciprocal Rank) and results merging algorithms (SSL, CORI). The CORI results merging algorithm (J. P. Callan et al., 1995) is based on a heuristic weighted scores merging algorithm. Semi-supervised learning (Luo Si & Callan, 2003a), makes use of a centralized index, comprised of all the sampled documents from the remote collections. The algorithm takes advantage of the common documents between the centralized index and the remote collections and their corresponding relevancy scores to estimate a linear regression model between the two scores.

We also tested different combinations of the number of collections requested and documents retrieved from each collection. Based on observations that relevant patents usually are located in few IPC main groups, subclasses or even subgroups we run experiments selecting 10, 20 or 50 sub-collections (IPC codes) requesting 100, 50 or 20 documents from each sub-collection respectively. In total we tested 44 combinations of DIR methods. We also performed two runs with the centralized index one with the spClms set of topics and the other with the AllClms set of topics. All the runs reported here are conducted with the Split-3 and Split-4 divisions. We didn't manage to perform any run with the Split-5 division due to increased time and other resources required.

4 Results

Table 3 shows the top 30 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 or split-4), method for source selection, method for merging, set of queries (spClms, AllClms) and, finally if regular expressions were used to remove numbers and words of less than two characters in the query (regexp).

Table 3. Results of the top 30 submitted and not submitted runs

Run description (runs denoted with an asterisk are one of the 8 submitted runs)	**PRES @100**	**MAP @100**	**Recall @100**
10-100.split4.CORI-CORI.AllClms	0,343	0,116	0,428
20-50.split3.CORI-CORI.AllClms	0,329	0,102	0,45
20-50.split4.CORI-CORI.AllClms	0,327	0,111	0,421
10-100.split4.CORI-SSL.AllClms	0,326	0,119	0,432
20-50.split4.CORI-SSL.AllClms.regex	0,325	0,127	0,439
50-20.split3.CORI-CORI.AllClms	0,318	0,09	0,441
20-50.split3.CORI-SSL.AllClms.regex	0,315	0,104	0,453
50-20.split3.CORI-SSL.AllClms.regex	0,311	0,105	0,444
50-20.split4.CORI-SSL.AllClms.regex	0,309	0,119	0,436
Centralized.Inquery.AllClms.regex	0,309	0,115	0,411
50-20.split4.CORI-CORI.AllClms	0,307	0,107	0,428
10-100.split3.CORI-CORI.AllClms	0,292	0,084	0,418
Centralized.Inquery.spClms.regex (*)	0,268	0,107	0,356
10-100.split3.CORI-SSL.AllClms	0,258	0,09	0,35
50-20.split3.CORI-CORI.spClms	0,248	0,094	0,294
20-50.split3.CORI-SSL.spClms.regex (*)	0,245	0,1	0,324
10-100.split3.CORI-CORI.spClms (*)	0,242	0,086	0,324
10-100.split3.CORI-SSL.spClms	0,236	0,09	0,326
50-20.split3.CORI-SSL.spClms.regex (*)	0,232	0,103	0,27
50-20.split4.CORI-SSL.spClms.regex (*)	0,23	0,103	0,275
20-50.split4.CORI-SSL.spClms.regex	0,213	0,904	0,27
10-100.split3.BordaFuse-SSL.AllClms.regex	0,211	0,042	0,311
50-20.split3.BordaFuse-SSL.AllClms.regex	0,211	0,04	0,32
10-100.split3.RR-SSL.AllClms.regex	0,202	0,04	0,311
50-20.split3.RR-SSL.AllClms.regex	0,202	0,04	0,32
20-50.split3.BordaFuse-SSL.AllClms.regex	0,199	0,036	0,289
20-50.split3.RR-SSL.AllClms.regex	0,199	0,035	0,289
10-100.split4.BordaFuse-SSL.AllClms.regex	0,195	0,037	0,237
50-20.split4.RR-SSL.AllClms.regex	0,195	0,037	0,237

5 Discussion

As it is shown in Table 3 the best performing combinations are those using CORI as the source selection algorithm and CORI or secondly SSL as the merging algorithm. The superiority of CORI as source selection method is unquestionable with the best run without CORI appearing in rank 21 (10-100.split3.BordaFuse-SSL.AllClms.regex). Also, it seems the best runs are those requesting fewer sub-collections 10 or 20 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 4). Table 4 reveals very 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 4. 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	4,49	2,41	4,51	1,87	1,99
EN ONLY	4,77	2,66	5,40	2,03	2,40
Split 4					
ALL	4,49	3,25	8,58	2,64	2,44
EN ONLY	4,77	3,54	9,77	2,76	2,86
Split 5					
ALL	4,48	8,43	24,33	2,89	4,90
EN ONLY	4,77	7,94	21,79	2,74	5,21

In general our approach has managed to score relatively high in the document level for which we submitted results. For example our top scoring approach (10-100.split4.CORI-CORI.AllClms, not included in the initially submitted 8 run files) has scored higher than any other submitted run.

Another interesting finding is that the queries that have been produced on the full set of claims that can be extracted from a topic produced significantly better and consistently better than the queries which have been produced from the specific claim or list of claims for each topic. This first approach (i.e. AllClms) can be considered as an implicit query expansion method which seems to have a positive effect on performance. Probably the SpcClms condition would produce better results for the passage retrieval task as the specific claim(s) probably contain more accurate information pertinent to support the task of retrieving a specific passage within a document, but as we focused only in the document level we couldn't test this reasonable hypothesis.

In addition to the comments already discussed, perhaps the most interesting and important finding for this study is that DIR approaches managed to perform better than the centralized index approaches, with 9 DIR combinations scoring better than the best centralized approach. This is a very interesting finding which 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 in patent domain the cluster-based approaches to information retrieval (Willett, 1988)(Fuhr, Lechtenfeld, Stein, & Gollub, 2011) 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 2012 passage retrieval starting from claims (patentability or novelty search) task. We tested different combinations of source selection (CORI, BordaFuse, Reciprocal Rank) and results merging algorithms (SSL, CORI). We have divided the CLEF-IP collection using the Subclass (Split-3) and the main group (Split-4) level to experiment with different levels and depth of topical organization.

We submitted 8 runs. According to PRES @100 our best DIR approach ranked 7th across 31 submitted results, however our best DIR (not submitted) run outperforms all submitted runs. The best performing source selection algorithm was CORI. On the other side the (precision oriented) methods Reciprocal Rank and BordaFuse consistently produced worse results.

We plan to explore further this line of work with exploring modifications to state-of-the-art DIR methods which didn't perform well enough in this set of experiments and to make them more effective for patent search. Also, we would like to experiment with larger distribution levels based on IPC (subgroup level). As we already explained we produced divisions of higher granularity at level 5 of IPC but we didn't have the time and the resources to report results for this division (split-5). We plan to report the runs using split-5 in a future paper.

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. In conclusion, we have illustrated that DIR methods could be 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 experiments presented in this paper.

ACKNOWLEDGMENT. The first 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), 22–29.
2. Aslam, J. A., & Montague, M. (2001). Models for metasearch. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 01* (pp. 276–284). ACM Press. doi:10.1145/383952.384007
3. Callan, J., & Connell, M. (2001). Query-based sampling of text databases. *ACM Transactions on Information Systems*, *19*(2), 97–130. doi:10.1145/382979.383040
4. Callan, J. P., Lu, Z., & Croft, W. B. (1995). Searching distributed collections with inference networks. *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 21–28). ACM. doi:10.1145/215206.215328
5. Chen, Y.-L., & Chiu, Y.-T. (2011). An IPC-based vector space model for patent retrieval. *Information Processing & Management*, *47*(3), 309–322. doi:10.1016/j.ipm.2010.06.001
6. French, J. C., Powell, A. L., Callan, J., Viles, C. L., Emmitt, T., Prey, K. J., & Mou, Y. (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*, 238–245. doi:10.1145/312624.312684
7. Fuhr, N., Lechtenfeld, M., Stein, B., & Gollub, T. (2011). The optimum clustering framework: implementing the cluster hypothesis. *Information Retrieval*, *15*(2), 93–115. doi:10.1007/s10791-011-9173-9
8. Larkey, L. S., Connell, M. E., & Callan, J. (2000). Collection selection and results merging with topically organized U.S. patents and TREC data. *Proceedings of the ninth international conference on Information and knowledge management - CIKM '00* (pp. 282–289). New York, New York, USA: ACM Press. doi:10.1145/354756.354830
9. Lupu, M., Mayer, K., Tait, J., & Trippe, A. J. (2011). *Current Challenges in Patent Information Retrieval*. (M. Lupu, K. Mayer, J. Tait, & A. J. Trippe, Eds.)*Information Retrieval* (Vol. 29, p. 417). Springer. doi:10.1007/978-3-642-19231-9
10. Miller, J. (2007). Most fed data is un-googleable. *Federal ComputerWeek*.

11. 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), 1580–1599. doi:10.1016/j.ipm.2007.12.008
12. Paltoglou, Georgios, Salampasis, M., & Satratzemi, M. (2009). *Advances in Information Retrieval*. (M. Boughanem, C. Berrut, J. Mothe, & C. Soule-Dupuy, Eds.) (Vol. 5478, pp. 497–508). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-00958-7
13. Paltoglou, Georgios, Salampasis, M., & Satratzemi, M. (2011). Modeling information sources as integrals for effective and efficient source selection. *Information Processing & Management*, *47*(1), 18–36. doi:10.1016/j.ipm.2010.02.004
14. Raghavan, S., & Garcia-Molina, H. (2001). Crawling the Hidden Web. *Proceedings of the International Conference on Very Large Data Bases* (Vol. 251, pp. 129–138). Citeseer.
15. Shokouhi, M. (2007). Central-Rank-Based Collection Selection in Uncooperative Distributed Information Retrieval. (G. Amati, C. Carpineto, & G. Romano, Eds.)*Advances in Information Retrieval*, *4425*, 160–172. doi:10.1007/978-3-540-71496-5_17
16. Si, L, Jin, R., Callan, J. P., & Ogilvie, P. (2002). A language modeling framework for resource selection and results merging. *ACM CIKM 02* (pp. 391–397). ACM Press. doi:10.1145/584792.584856
17. Si, Luo, & Callan, J. (2003a). A semisupervised learning method to merge search engine results. *ACM Transactions on Information Systems*, *21*(4), 457–491. doi:10.1145/944012.944017
18. Si, Luo, & Callan, J. (2003b). 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*, 298. doi:10.1145/860435.860490
19. Van Rijsbergen, C. J. (1979). *Information Retrieval*. *Information Retrieval* (Vol. 30, p. 208). Butterworths. doi:10.1016/0020-0271(68)90016-8
20. Willett, P. (1988). Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management*, *24*(5), 577–597. doi:10.1016/0306-4573(88)90027-1