Inf Retrieval J (2015) 18:559–585
DOI 10.1007/s10791-015-9270-2

Multilayer source selection as a tool for supporting patent search and classification

Anastasia Giachanou[1] · Michail Salampasis[2] · Georgios Paltoglou[3]

Received: 3 May 2013 / Accepted: 26 September 2015 / Published online: 27 October 2015

Abstract In this paper we present a method that can be used to attain specific objectives in a typical prior art search process. The objectives are first to assist patent searchers in understanding the underlying technical concepts of a patent by identifying relevant international patent classification (IPC) codes and second to help them conduct a *filtered* search based on automatically selected IPCs. We view the automated selection of IPCs as a collection selection problem from the domain of distributed information retrieval (DIR) that can be addressed using existing DIR methods, which we extend and adapt for the patent domain. Our work exploits the intellectually assigned classifications codes that are used to categorize patents and to facilitate patent searches. In our method, manually assigned IPC codes of patent documents are used to cluster, distribute and index patents through hundreds or thousands of sub-collections. We propose a new multilayer collection selection method that effectively suggests classification codes exploiting the hierarchical classification schemes such as IPC/CPC. The new method in addition to utilizing the topical relevance of IPCs at a particular level of interest exploits the topical relevance of their ancestors in the IPC hierarchy and aggregates those multiple estimations of relevance to a single estimation. Experimental results on the CLEF-IP 2011 dataset show that the proposed approach outperforms state-of-art methods from the DIR domain not only in

✉ Michail Salampasis
msa@it.teithe.gr

Anastasia Giachanou
anastasia.giachanou@usi.ch

Georgios Paltoglou
g.paltoglou@wlv.ac.uk

[1] Faculty of Informatics, Università della Svizzera Italiana (USI), Via Giuseppe Buffi 13, Lugano CH-6904, Switzerland

[2] Department of Informatics, Alexander Technological Educational Institute of Thessaloniki, 57400 Sindos, Greece

[3] School of Technology, University of Wolverhampton, City Campus, Office MI140, Wulfruna Street, Wolverhampton WV1 1LY, UK

identifying relevant IPC codes but also in retrieving relevant patent documents given a patent query.

Keywords Patent search · Patent classification · IPC · IPC suggestion · Distributed information retrieval

1 Introduction

Search technologies have been used for professional search (e.g. patent, medical, scientific literature search) for more than 40 years now as an important method for information access (Adams 2010). Despite the tremendous success of web search technologies, there is a significant skepticism from professional searchers and a very conservative attitude towards adopting search methods, tools and technologies beyond the ones which dominate their domain (Krier and Zaccà 2002). A typical example is patent search where professional search experts typically use the Boolean search syntax and quite complex intellectual classification schemes (Dirnberger 2011). Of course there are good reasons for this, apart from the psychological inertia produced by every successful scientific method exercised for a long period of time (Loh et al. 2006). For example, a patent search professional often carries out search tasks for which high recall is important. Additionally, s/he would like to be able to reason about how the results have been produced, the effect of any query re-formulation action in getting a new set of results, or how a set of query submission actions can be easily and accurately reproduced. Classification schemes are heavily used because it is widely recognized that once the work of assigning patent documents into classification schemes is done, the search can be much more efficient and language independent.

However, despite the overall skepticism, search technologies are being used increasingly in the workplace as a result of the explosion of content becoming electronically available, and those who deal with patents in their professional life are becoming more knowledgeable about new search technologies and tools (Atkinson 2008). As a result, even among the professionals, there is a feeling that intelligent search tools "that no longer just do what you say but also what you mean" (Wolter 2012) can and must help and this feeling is reflected in various studies (e.g. Bonino et al. 2010) and evaluation campaigns (e.g. Roda et al. 2009; Lupu 2011). Of course, adopting new search tools that will gradually change the landscape of patent search will be a long-term process. To achieve this, careful consideration is needed when designing a patent search system about the tools that will add functionality to ease the use of the core retrieval system but with the less invasive way in changing the traditional approaches of patent search.

The tools which will be integrated into patent search systems to assist professional searchers, and the way which these search tools will become part of a search system, do not only have to do with existing technologies, but probably more with the context and the attitude that a patent search is conducted. Furthermore, it is also very important to understand a search process and how a specific tool can attain a specific objective of this process and therefore increase its efficiency. For example, Lupu and Hanbury (2013) in a recent review of patent retrieval present a typical prior art search use case. The use case is analyzed in different sub-processes, performed by a patent examiner (pp. 15) to model and

better understand prior art search. This type of search is probably the most common patent search type.

In this paper, based on the context provided by the discussion above and leaving aside the core search engine, we discuss, improve and experiment with distributed information retrieval methods to set the ground for improving patent search. The improvement refers to the very fundamental step in professional patent search (sub-process 3 in the use case presented by Lupu and Hanbury which is "defining a text query, potentially by Boolean operators and specific field filters"). In prior art search probably the most important filter is based on the International Patent Classification (IPC or CPC)[1] classification (Vijvers 1990; Adams 2000). Selecting the most promising/relevant IPCs depends of course on the prior knowledge of a patent professional in the technical area under examination, but sometimes the area of a patent application may not be easily distinguishable or usually a patent uses various technical concepts represented by multiple IPCs. To identify all these relevant IPCs could be a difficult, error prone and time-consuming task, especially for a patent professional not very knowledgeable in some technical area.[2]

The method which we present in this paper can be the basis for a search tool to support this step automatically; this is, given a query, select the most appropriate IPCs, make a filtered search based on the automatically selected IPCs, and finally merge the results returned from each suggested IPC (i.e. an IPC code is seen as a DIR sub-collection in our method). This process very precisely and naturally resembles the way patent professionals conduct various types of patent searches. Also, the patent searcher may use the tool not only to produce IPC-based filters to automatically narrow a search, but also as a classification search which will be used as a starting point to identify and closer examine technical concepts as these are expressed in IPCs and to which a patent could be related and should be examined more vigorously. This ground understanding step helps soon after in formulating better queries with higher precision, which will usually include expansion with noun-phrases from the IPCs, which deemed relevant.

The IPC suggestion task as we explore it in this paper can be seen as an instance of the patent classification problem in the sense that it aims to identify IPC codes as it happens in patent classification. The only difference between the two tasks has to do with the different objectives they try to attain after IPC selection/patent classification. IPC suggestion aims to identify IPC codes that contain relevant patent documents and in this way supports a narrower filtered search while in the case of the patent classification a patent document is assigned one or more IPC classification codes. Due to the similarity of the two tasks, we further discuss the literature focused on the patent classification in Sect. 2.4.

The rest of this paper is organized as follows. In Sect. 2 we present in detail how patents are topically organized using their IPC code, how our method exploits this hierarchical taxonomy, the DIR technologies, which we use and extend and we discuss previous work focused on patent classification. In Sect. 3 we present a new methodology for collection selection proposed in this paper. This new method relies on using multiple evidence taken from various levels of IPC. In Sect. 4 we describe the details of our experimental setup and

[1] CPC is the new official classification scheme endorsed by EPO and USPTO but the test collection that we used for this study (CLEF-IP) has only IPC codes therefore we use the term IPC. However the results are equally transferable to CPC and generally speaking to any hierarchical classification scheme based on the same principles as IPC.

[2] In smaller patent offices without many patent examiners, they usually cover and are asked to examine patents in broad technical areas.

the experiments conducted. We continue with the results and a discussion of our approach in Sect. 5 and future work and conclusions in Sect. 6.

2 Distributed IR and topically organized patents

2.1 Distributed information retrieval

Distributed Information Retrieval (DIR), also known as federated search (Si and 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 and 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 are combined into a single merged result list which is returned to the user (Si and Callan 2003a; Paltoglou et al. 2008).

DIR has been explored for about 20 years now.[3] One recent application of DIR methods is the aggregated or vertical web search (Arguello et al. 2009). Also many enterprise search applications rely on forms of DIR. Additionally it is known that big web search companies use, at physical level, DIR techniques in maintaining distributed indexes mainly for scalability reasons, however at a logical level their search services are perceived as centralized by the end-users. Another widespread application of DIR methods is the digital library search systems (Buckland and Plaunt 1997; Larson 2003).

Generally speaking, DIR had several successes and applications; however, it is not as much widespread as the centralized approaches in most traditional search applications. The main reason is probably that distributed information retrieval approaches manage to perform similar or better than the centralized approaches only in cases where the data are well organized but this is rarely possible (Powell et al. 2000). One recent attempt to promote research in federated search was the launch of the Federated Web Search evaluation track (Demeester et al. 2013). The aim of the track was to encourage researchers to develop approaches that can be used in a realistic setting and can tackle the problem of the federated search. In 2013, the track was focused on both resource selection and results merging tasks.

Although there is no clear distinction between the terms DIR and federated search (Shokouhi and Si 2011), the latter represents a more realistic DIR scenario that allows the simultaneous search of multiple searchable, remote and *physically* distributed, resources. ezDL is probably the most representative interactive search system and general framework using federated search techniques (Beckers et al. 2012). However, in accessing hidden web resources using a federated search system, there are important challenges and problems, which need to be solved. These challenges include managing credentials for accessing search systems residing behind turnstiles, maintaining wrappers which may need to be frequently updated due to changes in the web resources accessed, translating the federated system's query format into the format accepted by remote resources, and finally mapping the results of multiple remote sources into a common central format.

[3] The pioneering work made by Voorhees et al. (1994) and Moffat and Zobel (1994) can be regarded as the first steps in DIR.

In this paper we try to reconsider DIR methods in patent search on the basis of a different motivation from that of the original DIR research. We believe this approach is original and interesting because we use it to attain different objectives. Furthermore, we believe that our approach can be widely applicable, because in professional search quite often documents are hierarchically organized into thematically coherent collections. Patent search is a very good example because patents have intellectually assigned classification codes providing an organized environment where DIR techniques can be effectively applied. In our study, the IPC codes existing in every patent are used to topically cluster, distribute and index patents through hundreds or thousands of sub-collections. Our method and the tool that can be developed upon it automatically selects the best sub-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.

The work that is presented here is different from typical DIR studies because we assume a cooperative environment wherein the collections' statistics can be easily accessed. Cooperative environments provide important statistics about the collections' contents in contrast to the uncooperative that do not provide details about their contents and thus require query-based sampling. We can assume such environment because unlike environments with an unknown and rapidly growing number of not directly accessible documents (i.e. the deep web), the patent domain contains a certain number of documents and a single point of authority can be established as it happens in a cooperative environment. Assuming a cooperative environment, we did not apply any source sampling before the source selection, a phase which is required in uncooperative environments. But most importantly our work is not a typical web-based DIR work because our main target is source selection (i.e. selecting the most relevant IPCs) rather than producing better patent retrieval results which largely relies on effective merging, although this is also possible.

Additionally, this work is not a typical federated search study, since we focus more on logically clustering the patents rather than distributing them at a physical level. We create clusters of patents based on their manually assigned IPC codes and we test different collection selection methods for the IPC selection task. In that sense our work could be seen as a method for cluster-based document retrieval using DIR selection methods. However, it differs from cluster-based retrieval because our main goal is not to retrieve clusters but rather to identify sub-collections which will be later searched more thoroughly using fielded (filtered) search. Quite naturally then, in our work we didn't give any attention 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 and the patent collections can be indexed centrally, probably this might not be the case, depending on the communication and other costs for crawling patent documents and building the centralized index. On the other hand, it should be equally said that because our method is based on logically distributing the patent documents and not physically distributing them, the DIR methods which we applied (for source selection but also for results merging) can also operate very efficiently.

In this paper, we extend our previous work of applying DIR methods to topically organized patents (Salampasis et al. 2012). 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. We propose a new collection selection method that surpasses previous source/IPC selection methods for topically organized patents. In the experiments we used both the standard CORI and ReDDE algorithms and the fusion-based source selection methods (Paltoglou et al. 2009) using Reciprocal Rank and BordaFuse (Aslam and Montague 2001) for comparison.

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. This dataset was significantly smaller and they used few queries (37) in comparison to our study (300). But most importantly the division into sub-collections in our work happens into much larger scale. They divided patents into 401 collections while our method exploiting the hierarchical structure of IPC uses significant larger divisions (e.g. 63.806 collections at level 5). Field studies and existing tools (e.g. Espacenet Classification Search) show that our approach of dividing patents is closer to the actual way of patent professionals conducting patent searches, as they usually conduct classification searches at level 4 or 5. Additionally, our approach to apply a standard source selection algorithm in multiple layers is new and much more effective because it supports finding relevant IPCs at different levels of the IPC hierarchy.

A significant number of methods have been proposed with the aim to improve the prior-art search by utilizing the IPC codes assigned to the patent documents and the patent applications/topics (Itoh 2005; Konishi 2005; Harris et al. 2011; Cetintas and Si 2012). The majority of those methods utilize the IPC codes that are already assigned to the patent application/topic that is under examination in order to improve the patent search. Some researchers excluded the retrieved patent documents that did not partially match the IPC codes of the patent topic (Itoh 2005) while others used methods that boost patents documents that share common IPCs with the patent topic (Konishi 2005). Other researchers added more sophisticated features that also require the IPC codes of the patent application to be known before the prior-art search (Cetintas and Si 2012). However, the assumption that the IPC codes of the topic are known before the prior-art search is not usually a realistic scenario. Our work is different, as our method utilizes the IPC codes attributed to the patent documents and not those attributed to the patent topic. Therefore, we believe that we consider a more realistic environment wherein it is not required for the patent examiner to know the topic's IPC codes before the prior-art search.

Our work can be considered also relevant to a recent DIR work which aims to reduce the uncertainty in resource selection (Markov et al. 2013). Similar to this work (although not so extensively) we obtain a number of estimates for source selection, rather than relying upon only one point estimate. But what is more important in our work is that we reconsider DIR methods in patent search on the basis of a different motivation. We try to transparently assist professional searchers, which are very reluctant in accepting black box approaches. Our emphasis and priority on source selection for presenting relevant IPCs to professional searchers instead of focusing on merging and presenting a single merged result aims to create this feeling of transparency and control to the end user.

2.2 Prior work on collection selection

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 and Connell 2001).

The collection retrieval inference network (CORI) algorithm (Callan et al. 1995) is one of the most widely used source selection algorithms. The algorithm creates a hyper-document for each sub-collection, containing all the documents that are members of the sub-collection. When a query Q is submitted, the sub-collections are ranked based on the

belief $p(Q|C_i)$ the collection C_i can satisfy the information need of the query Q. The belief $p(r_k|C_i)$ that a term r_k of the query Q, is observed given collection C_i is estimated as:

$$T = \frac{df}{df + 50 + 150 * \frac{cw}{avg\_cw}}$$

$$I = \frac{\log\left(\frac{|N|+0.5}{cf}\right)}{|N| + 1.0}$$

$$p(r_k|C_i) = b + (1 - b) * T * I$$

where *df* is the number of documents in collection C_i that contain term r_k, cf is the number of collections that contain term r_k, cw is the number of terms in C_i, avg_cw is the average cw, |N| is the number of available collections and b is the default belief, set to the default value of 0.4. The overall belief $p(Q|C_i)$ in collection C_i for query Q is estimated as the average of the individual beliefs of the representation concepts: $p(Q|C_i)$

$$p(Q|C_i) = \frac{\sum_{r_k \in Q} p(r_k|C_i)}{|Q|}$$

The Decision-Theoretic framework (DTF) presented by 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 relevance of each remote collection to explicitly estimating the number of relevant documents in each. ReDDE (Si and Callan 2003b) focuses on 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 when querying the centralized sample index. Its performance has been shown to be 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 (Paltoglou et al. 2009). Different voting mechanism can be used (e.g. BordaFuse, ReciRank, Compsum) mainly inspired by data fusion techniques. In data fusion techniques, when the user submits a query to the system, the first step is to produce a ranking of retrieved documents from the centralized index. Let $R(Q) = \{D_1, D_2, D_3, \ldots, D_n\}$ be the set of documents retrieved for query Q. A ranking of collections for query Q can be produced by calculating a score for each collection (C) as an aggregation of votes from all documents D_i that are retrieved from the centralized index. This set is referred to as Votes(C,Q). Both BordaFuse (Aslam and Montague 2001) and ReciRank utilize the rankings of the documents to calculate the score for each collection. The following equations show respectively the score of collection C for query Q as calculated according to BordaFuse and ReciRank:

$$score\_bordafuse(C,Q) = \sum_{D_i \in Votes(C,Q)} (|R(Q)| - rank_{D_i})$$

$$score\_RR(C,Q) = \sum_{D_i \in Votes(C,Q)} \frac{1}{rank_{D_i}}$$

There is a major difference between CORI and the other source selection algorithms tested. CORI builds a hyper-document for each sub-collection while the other collection selection methods are based on the retrieval of individual documents from the centralized sample index. Due to its main characteristic CORI has been repeatedly reported in the literature (Powell and French 2003) not performing consistently well in environments containing a mix of "small" and "very large" document collections.

However, in the patent domain where similar inventions contain to a large extent very different terminology (Larkey 1999) the idea of building hyper-documents centered around a specific technical concept such as IPCs is well suited. The homogenous collections containing patent documents of the same IPC as the hyper-documents in CORI should normally encompass a strong discriminating power, something very useful for effective and robust collection selection.

2.3 Topically organized patents

All patents have manually assigned IPC codes (Chen and Chiu 2011). IPC is an internationally accepted standard taxonomy for classifying, sorting, organizing, disseminating, and searching patents and is officially administered by the 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 codes, 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 that is a prominent prerequisite for DIR, there are some important differences, which motivated us to re-examine and adapt existing DIR techniques as they are applied in patent search and in the context provided by our specific objectives. First, IPC codes 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 (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 not missing relevant patent documents given a specific topic is of high importance for the patent professionals. For example, a single missed patent in a patentability search that could invalidate a newly granted patent can lead to lawsuits due to patent infringement. This contrasts with web search where high precision of initially returned results is the requirement and on which DIR algorithms were mostly concentrated and evaluated (Paltoglou et al. 2008).

Before we describe our study further we should explain more the IPC scheme, which determines how we created the sub-collections in our experiments. Top-level IPC codes 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. Each subclass is divided into main groups, which are further subdivided into subgroups. In total, the current IPC has 8 sections, 129 classes, 632 subclasses, 7530 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 that 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. at the IPC hierarchy). Hierarchies among subgroups (i.e. below level 5) are indicated not by subgroup symbols but by the number of dot symbols preceding the node descriptions as shown in Table 1.

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. Since 2010, the IPC is revised once a year. Sometimes existing subclasses/groups/subgroups are subdivided into new subsets. Sometimes a set of subclasses of a class are merged together, and 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. The changes are related to a small part of the IPC hierarchy and therefore they are negligible.

In the experiments presented in Sect. 4 we divided the CLEF-IP collection using the subclass (split3),[4] 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 narrower sub-collections. In the experiments reported here, 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. For example, if one patent is assigned the following IPC codes {F28D15/04, G11B20/02, F28D15/02, G06F17/30} then this patent belongs to three subclasses represented by sub-collections in split3 {F28D, G11B, G06F}, three main groups represented by sub-collections in split4 {F28D15, G11B20, G06F17} and four subgroups {F28D15/04, G11B20/02, F28D15/02,

[4] In this paper, we use the term "level" to refer to the different levels in the IPC hierarchy while the term "split" refers to the different collections that were created using the levels of the IPC hierarchy. The term "split" is used to represent the settings wherein we conducted our experiments.

Table 1 An example of a section from the IPC classification

Division	Title	IPC code
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

G06F17/30}. This is the reason why the column #patents (Table 2) presents a number larger than the 1.3 million patents that constitute the CLEF-IP 2011 collection and also why the number of patents is different at each split. Due to this fact we had to make some modifications to existing DIR algorithms since most of them assume not overlapping sub-collections.

From the statistics presented in Table 2, we can observe that there are sub-collections that contain a single document. This is a result of large number of IPC classification codes and the necessity to use an evaluation track for our experiments so we can produce comparable and reproducible results. Considering the large number of the classification codes, it was likely that some sub-collections would have one single document. In a real environment, we believe that there are classes, especially at the level of subgroup, with few patents considering the granularity of the hierarchy and the large number of the classification codes. In order to avoid the bias in the results, we do not exclude those collections from the experiments.

2.4 Relationship to patent classification

The problem of automated IPC suggestion can be viewed as a large scale text classification problem, which refers to the task of classifying a document to one or multiple classes/categories (Kosmopoulos et al. 2010). More recently hierarchies have become more popular for the organization of text documents. This is mainly because many real world systems use taxonomies to better classify and organize data over a set of hierarchically organized categories with parent–child relations. Web directories and Wikipedia are two examples of such hierarchies. In the patent domain, where IPC has been used as the hierarchical taxonomy for many decades now for organizing patents, the main challenges investigated from the IR and the machine learning communities[5] have to do with the very large hierarchies (e.g. IPC at level 5 has more than 70.000 IPCs/categories) and the dynamic nature of the IPC scheme as it is periodically expanded by adding new categories, thus requiring reclassification of existing patents. Before the rise of patent classification as a machine-guided task, most text classification evaluation tasks covered a smaller number of documents. For example, there are 103 categories in the Reuters framework (Lewis et al. 2004).

More specifically, the IPC suggestion task as we explore it in this paper can be also viewed as a patent classification problem in the sense that it aims to identify IPC codes

[5] Challenges on Large Scale Hierarchical Text classification: http://lshtc.iit.demokritos.gr/.

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

Split	No. of patents	Number of IPCs (sub-collections)	Docs per collection			
			Avg	Min	Max	Median
split3	3622570	632	5732	1	165,434	1930
split4	5363045	7530	712	1	83,646	144
split5	10393924	63,806	163	1	39,108	36

rather than documents given a patent query. Since the late 1990s automating the process of patent classification has received much academic attention and many researchers tried to address this task by following a number of different techniques such as modifying and extending a conventional text classification algorithm in the context of the patent domain (Larkey 1998; Kohonen et al. 2000; Fall et al. 2003; D'hondt et al. 2013), incorporating the hierarchy into the classification algorithm (Chakrabarti et al. 1998; Cai and Hofmann 2004; Tikk et al. 2007) or using linguistic analysis (Gey et al. 2001; D'hondt et al. 2013). A comprehensive introduction to the task of automated patent classification is given by Benzineb and Guyot (2011).

Aiming to promote research in patent classification, Krier and Zaccà (2002) organized a comparative study of various academic and commercial patent classification systems. The best results were achieved by Koster et al. (2001) who used the Balanced Winnow algorithm. In their study, categorization was performed for 44 or 549 IPC categories and they achieved 78 and 68 % precision respectively at 100 % recall. In recent years, CLEF-IP has organized patent classification evaluation tasks and has provided to the researchers very large patent data to train their systems on more realistic data sets (Piroi et al. 2010, 2011, 2012). A number of approaches were proposed to address the patent classification using the CLEF-IP data sets (Guyot et al. 2010; Derieux et al. 2010; Verberne and D'hondt 2011; Beney 2010). In the CLEF-IP 2010 classification track, Guyot et al. (2010) achieved the best results using the Balanced Winnow algorithm. The categorization was performed on the subclass level and was based on a combination of words and statistical phrases.

Our approach is more related to the approaches that have tried to address the patent classification by incorporating information from the hierarchy into the algorithm. One of the first attempts to incorporate the hierarchy into the categorization algorithm was made by Chakrabarti et al. (1998) who performed some small-scale tests based on a Bayesian hierarchical patent classification system which could classify the patents into 12 subclasses organized in three levels. This work demonstrated that using the classifications of cited patents could improve the categorization of the patents. Cai and Hofmann (2004) proposed a multiclass SVM for hierarchical categorization. In their work the loss functions were redefined. The minimum accuracy they obtained at the IPC main group level was 32.4 % at section F while the maximum accuracy obtained at section A was 42.9 %. Tikk et al. (2007) presented a hierarchical online classifier called HITEC algorithm which applies a neural network with the aim to utilize the hierarchical structure of patent taxonomy.

Most of the methods that have been proposed and applied on patent classification were evaluated at the subclass and main group level and the classification was restricted to a rather small number of classes. In contrast, our method can be effectively applied at the level of subgroup that has about 70,000 codes. One of the works focused on the subgroup level was presented by D'hondt, E. (2014) who proposed adding phrasal features to

unigram features to improve classification at subgroup level. Both WIPO-alpha collection containing 75,250 patent documents and a subset of CLEF-IP with 991,805 patent documents were used to evaluate the proposed approach. The results showed that combining a two-step hierarchical approach with unigram and skipgram features obtained 32.5 % F1 accuracy for the WIPO-alpha set and 38.4 %, on the subset of CLEF-IP 2010 on subgroup level classification. However, since skipgrams are sparse, the approach needs a sufficiently large amount of training data. In addition, Chen and Chang (2012) proposed a three phased categorization method that obtained an accuracy of 36.07 % at the level of subgroup. However, the effectiveness of their approach is examined on a much smaller collection as they used the English WIPO-alpha collection containing 75,250 patent documents.

3 Multilayer collection selection

We exploit the aforementioned hierarchical organization of the IPC classification scheme and the ideas of (a) considering topically organized patents using IPC as a DIR system where IPC codes act as sub-collection identifiers with parent–child relations and (b) creating a more effective and reliable source selection method relying upon a weighted sum of multiple estimates, to propose a new *multilayer* collection selection method. The new method in addition to utilizing the topical relevance of collections/IPCs[6] at a particular level of interest exploits the topical relevance of their ancestors in the IPC hierarchy and aggregates those multiple estimations of relevance to a single estimation. For example, assume that the patent expert is interested in distilling the most relevant IPC codes at level 5 (i.e., subgroup level—refer to Table 1) for a prior-art search. The proposed methodology will combine the relevance of collections at level 5 (the level of interest) *and* level 4 (the ancestral level to the one of interest) in order to produce the final ranking. The method can be applied in any domain where documents are organized in accordance to a hierarchical classification scheme, but we focus here on the patent domain. In addition, because the method will exploit for each level the relevance of its ancestor level, recursively and overall the multilayer method exploits multiple levels of hierarchy (e.g., levels 4, 3, 2, 1 in the aforementioned example). However, exploitation of higher level of the hierarchy (e.g. level 1 in IPC) where very few classes exist may not be useful. In this work we limit the approach to only one level and leave the exploration of using multiple levels at a higher level as future work.

The algorithm functions in the following manner. Given a query document P and a target level i (for example if retrieval of IPCs at the level of subgroup is required, then the target level is 5) the algorithm produces a ranking $R_i(P) = \{C_1^i, C_2^i, \ldots, C_m^i\}$ *and scores* using a source selection algorithm and this can be formulated as:

$$\begin{aligned} R_i(P) &= \{C_1^i, C_2^i, \ldots, C_m^i\} \\ Score_i(P) &= \{Score(C_1^i), Score(C_2^i), \ldots, Score(C_m^i)\} \end{aligned} \tag{1}$$

Then given that the level of interest of the patent expert is i, the second phase of the algorithm is to re-rank $R_i(P)$ by utilizing the other estimates that will be produced from the ancestor level. So the second produced ranking is at the ancestral level $i-1$ of the

[6] In the remainder of the paper we utilize the terms "collection" and "IPC code" interchangeably, that is, when we are referring to a collection, we are implicitly referring to all the patents that have the same IPC code for a particular level (i.e., all the patents that belong to the F01 class, are part of the F01 collection at level 2).

hierarchical classification scheme. We symbolize this ranking as: $R_{i-1}(P)$, where C_j^{i-1} is the collection retrieved at rank j at level $i-1$ and n is the total number of collections retrieved at this level. The ranking and relevance scores are produced by applying any standard source selection algorithm at collections of this level and can be formulated as:

$$\begin{aligned} R_{i-1}(P) &= \left\{C_1^{i-1}, C_2^{i-1}, \ldots, C_n^{i-1}\right\} \\ Score_{i-1}(P) &= \left\{Score\left(C_1^{i-1}\right), Score\left(C_2^{i-1}\right), \ldots, Score\left(C_n^{i-1}\right)\right\} \end{aligned} \tag{2}$$

In the experiments described in this paper we use the CORI collection selection algorithm as it has been shown to be more effective than other collection selection algorithms, such as BordaFuse, ReciRank, ReDEE that were tested before (Salampasis et al. 2012).

After the calculation of $R_{i-1}(P)$ a re-ranking process is launched at the target level $R_i(P)$, by calculating a weighted sum, which includes the R_i(P) ranking, but it also takes into account the $R_{i-1}(P)$ ranking in the following manner: For each collection C_j^{i-1} *in* $R_{i-1}(P)$ previously retrieved, we locate its children in the IPC hierarchy: $Children(C_j^{i-1}) = \{C_j^i, where\ C_j^i\ is\ a\ child\ of\ C_j^{i-1}\}$. Subsequently, to re-rank $R_i(P)$ we locate the $Children(C_j^{i-1})$ for all C_j^{i-1} that belong to $R_{i-1}(P)$ and we re-rank $R_i(P)$ by recalculating the final relevance score of collection C_j^i at the level of interest i as follows:

$$FinalScore\left(C_j^i\right) = (1-a) * Score\left(C_j^i\right) + a * Score\left(C_j^{i-1}\right), \quad \exists! C_j^{i-1} \quad where \quad C_j^i \in Children\left(C_j^{i-1}\right) \tag{3}$$

where $Score(C_j^i)$ is the relevance score of collection C_j^i using any source selection algorithm (as previously mentioned, CORI in this case), $Score(C_j^{i-1})$ is the relevance score of collection C_j^{i-1} (the ancestral collection of C_j^i) using CORI, and a is the mixturing parameter that determines the weight that each level will have.

A parameter of our method is the *collection window m* (present in formula 1) which represents the number of sub-collections that will be considered for re-ranking from R_i(-P) after taking additional evidence from $R_{i-1}(P)$. For example if the aim is to produce a list of N suggested IPCs at level 5 to the end user, the method should define how many IPCs from the initial rank, and initially positioned after position N (note that N $\leq$ m), will be reconsidered in the re-ranking process. This *window* parameter m can be set either to a fixed threshold such as 100 or to a number relative to the number of IPCs that should be suggested (i.e. 2 * N, 3 * N etc.).

Another important parameter of our method is the *influence factor n* (present in formula 2), that is, how many IPCs/sub-collections from $R_{i-1}(P)$ will be utilized to *re-rank* IPCs found *inside the collection window* in the target level of interest. For example, if we want to return k IPCs at the level of interest (e.g., level 5), a parameter in our method is how many IPCs from level 4 will be used as additional evidence to affect the final ranking that will be produced applying formula 3 during the re-ranking process. An interesting observation is that the influence factor parameter, in broad terms, can be seen as the k parameter in the k-Nearest Neighbors algorithm (k-NN). k-NN introduced by Fix and Hodges (1951) is one of the simplest and most popular classification algorithms. In short, for text categorization the algorithm categorizes a new document using evidence from its k nearest neighbors. In broad terms the k can be viewed as the influence factor n in the context that it determines how many sub-collections from $R_{i-1}(P)$ will be considered for re-ranking $R_i(P)$.

For the experiments presented in this study, the value of the parameter α was decided after a training process and both the *collection window m* and *influence factor n* parameters

from previous experiments. During the training process that preceded the actual runs,[7] we tested various parameters to examine which values optimize the performance of the method. We used 300 English topics to train the multilayer algorithm. Potentially, in the future, we can optimize the parameters m and n for each individual query based on the decision-theoretic framework by Fuhr (1999), so that the expected overall cost of each retrieval process is minimized, in respect to the number of documents that are requested both at the IPC level of interest and the ancestral level.

As previously mentioned, we select CORI as the underlying source selection method because previous studies showed that it performs better than other collection selection methods (BordaFuse, Reciprocal Rank) when applied at the patent domain (Salampasis et al. 2012; Giachanou et al. 2013). We believe the reason this happens is because CORI represents collections using *hyper-documents*, while the other methods use documents retrieved from the centralized collection to estimate its relevance.

However, CORI tends to produce poorer results at lower IPC levels (e.g. from level 3 to level 4) that is logical since the problem is much more difficult (more classes, more similar classes). One reason is that the technological area of patents belonging to a collection is more accurately represented in higher IPC levels (e.g., subclass) because they consist of fewer IPC codes. At higher IPC levels, documents in one sub-collection are relatively homogeneous and better distinguished from those in other sub-collections, something that is more difficult to capture in lower levels of IPC hierarchy. For example, IPC codes at level 4 are less differentiated between each other using a hyper-document approach from those at level 3, resulting in a decreased CORI performance. It should also be noted that level 4 contains about ten times more sub-collections than level 3. 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.

Overall the multilayer method uses a re-ranking process, which takes into account the source selection results at several classification levels. Of course, as it has been already mentioned, the proposed method can recursively utilize multiple evidence in all levels of the classification scheme, but in this paper, we focus on level 3 (subclass), level 4 (main group) and level 5 (subgroup). From a conceptual point of view, score aggregations like the one used in our method have been proposed for other forms of structured retrieval (e.g. XML retrieval and hierarchical classification). Structured documents containing multiple metadata can be considered as having an internal hierarchical structure. On the basis of this structure, a number of researchers presented methods which function by aggregating evidence for relevance from different sources (Sigurbjörnsson et al. 2004; Kong and Lalmas 2007).

Finally, we should mention that our approach can be also considered relevant to classifier stacking approach, which is a type of ensemble learning. The main concept in stacking is that the output of the base classifier is used as training data for the next classifier. Similarly, we use evidence from the higher levels in order to identify the IPC codes at lower levels. However, in our approach we do not use classifiers from the field of machine learning, as they require a great amount of training data to do accurate predictions.

[7] The training runs were executed on a different set of queries than the one we used for the experiments below, that is, the set of training and testing queries are disjoint.

4 Experiments

4.1 Experimental setup

The dataset that is 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 and extended by documents from the WIPO. The patent documents have XML format and contain content in English, German or French (Piroi et al. 2011). The XML fields can be in more than one language. We indexed the collection with the Lemur toolkit.[8] The fields that 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 also involves stop-word removal using the Inquery's standard stop words list and stemming using the Porter stemmer (Porter 1980). In our study, we use the Inquery algorithm (Allan et al. 2000) as it is implemented in the Lemur toolkit.

To test our system, we use a subset of the official queries provided in CLEF-IP 2011 dataset. The topic collection contains 3973 query topics in English, German or French. The query topics are patents documents and have XML format. We run the first 300 English topics generated using the title, the abstract, the description and the claims. The topics represent a subset of the English topics because, as already mentioned, a different subset of the English topics was required to train the multilayer method. We tested different combinations of source selection (CORI, BordaFuse, ReciRank, ReDDE and multilayer) and results merging algorithms (SSL, CORI) at split3, split4 and split5. The CORI results merging algorithm (Callan et al. 1995) is based on a heuristic weighted scores merging algorithm. Semi-supervised learning (Si and Callan 2003a), makes use of a centralized index, which in our method was comprised of the whole set of documents from the dataset. The algorithm takes advantage of the common documents between the centralized index and the remote collections and their corresponding relevance scores to estimate a linear regression model between the two scores.

We also test different combinations of the number of collections requested and documents retrieved from each collection. Based on the observation that relevant patents are usually located in few IPC main groups (Table 3), subclasses or even subgroups, we run experiments selecting 10 or 20 sub-collections (IPC codes) for the source suggestion task (metric R_k). For the actual patent retrieval tasks (metric PRES, MAP and Recall) reported in Table 5 we retrieve 100 or 50 documents from each selected sub-collection respectively.

Multilayer algorithm required a training process during which we tested various parameters to examine which values optimize the performance of the method. For the training process, we use the succeeding 300 English topics after the topics used in the experiments. The same topics are used for the training of the three different splits. The multilayer algorithm is tested at split4 and split5. To produce IPC suggestions at IPC level 4 the multilayer used evidence from the collections retrieved at split3 and split4 while the IPC suggestions at split5 were produced using evidence from split4 and split5. To select IPC collections at split4 the *influence factor* parameter was set at 20 sub-collections while the *collection window* parameter at 200. At split5 the *influence factor* was set at 200 sub-collections and the *collection window* parameter at 2000 sub-collections.

We also performed a run with the centralized index to examine if our method can outperform the centralized approach, although this is not our primary objective. Lastly, we

[8] http://www.lemurproject.org/.

Table 3 Analysis of IPC distribution of topics and their relevant documents

IPC level	No. of topics	Average number of			
		Relevant docs per topic (a)	IPC classes of each topic (b)	IPC classes of relevant docs (c)	Common IPC classes between (b) and (c)
Training					
split3	300	8.22	2.08	4.8	1.76
split4	300	8.22	3.1	8.76	2.34
split5	300	8.22	5.82	19.84	3.63
Testing					
split3	300	8.57	2.09	5.15	1.75
split4	300	8.57	2.95	9.02	2.21
split5	300	8.57	5.58	20.56	3.73

performed runs with the optimal approach for each split to set an upper limit for DIR methods. For the optimal run, the system retrieved documents only from the collections containing the relevant documents.

In order to compare the performance of the collection selection methods we use the measure R_k proposed by French and Powell (2000). This measure has been widely adopted in the DIR research community (Callan et al. 1995; Nottelmann and Fuhr 2003; Si and Callan 2005) and is very important for evaluating collection selection algorithms at recall-oriented tasks as it provides an indication if an algorithm is able to rank sub-collections containing a large number of relevant documents high in the ranking. R_k compares a collection selection algorithm at rank n to the optimal collection selection algorithm at the same rank, which is ranking collections according to the number of relevant documents they contain for a query.

Additionally to R_k, we use the recall-oriented measures PRES and Recall to compare the performance of the various methods tested in their ability to retrieve patent documents. The Patent Retrieval Evaluation Score (PRES) is a metric appropriate for recall-oriented domains as it reflects the ability of the system to retrieve a large portion of the relevant documents in relatively high ranks given a user specific cut-off (Magdy and Jones 2010). Unlike PRES, the standard recall metric emphasizes on retrieving a large portion of the relevant documents without considering the rank. Generally PRES was preferred in patent retrieval evaluation tasks (Piroi and Zenz 2011; Piroi et al. 2012). Since patent professionals examine a large number of patents in order to find the relevant ones, we report PRES and recall up to 100 patent documents. Finally, to offer a more complete comparison of the multilayer algorithm, we also evaluated its performance using the MAP measurement.

5 Results and discussion

5.1 Parameter variation

As already mentioned, for the experiments presented in this study, a training process on the multilayer algorithm preceded the actual runs. To avoid over-fitting, we separated the topics into training and testing set. During the training process, we tested different values

Table 4 R_k for the training queries set

R_k @ 20 (training queries)									
Parameter a	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
split4 Influence factor: 20 Collection window: 200	0.653	0.655	0.657	0.667	0.668	0.673	0.674	0.677	0.676
split5 Influence factor: 200 Collection window: 2000	0.286	0.308	0.332	0.364	0.405	0.461	0.52	0.564	0.531

of the parameter to determine which one optimizes the performance of the method. We trained the method for split4 and split5. Table 4 summarizes the results from applying different parameter values in terms of R_k.

The decision on the values of the *influence factor* and *collection window* parameters was based on previous experiments in which we followed one round validation. To this end, we used half of the English topics that were used in this study, to optimize the parameters and half of them to test the methods. The influence factor and the collection window parameters were optimized using the following methodology. First, the influence factor was set to a value of 1.5 % of the number of sub-collections. Then the number of sub-collections was repeatedly increased by a step of 1.5 % until an influence factor of 50 %. We measured the PRES value for each different parameter value. The similar process was also followed for the collection window.

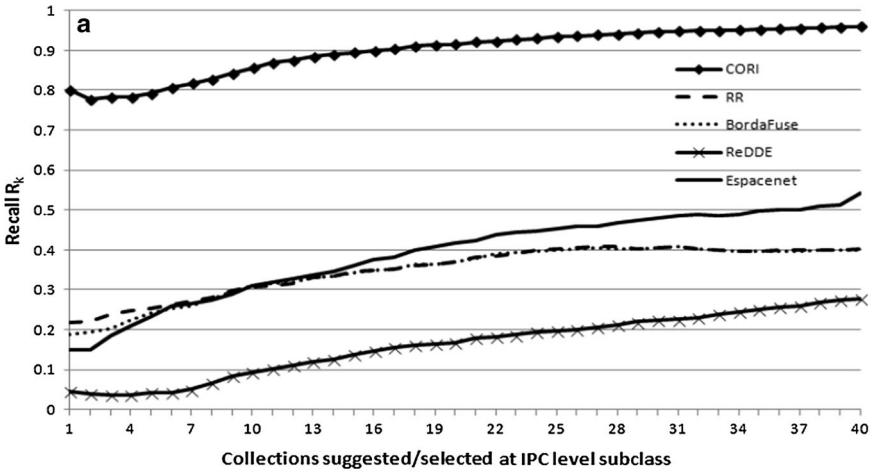
The *influence factor* parameter when retrieving from IPC level 4 was optimized at 20 sub-collections and the *collection window* parameter at 200 sub-collections. The same parameters when running the experiment at level 5 were optimized respectively at 200 sub-collections and 2000 sub-collections. The selection of the values of the influence factor and the collection window parameters can also be related with the basic structure of the IPC scheme. More specifically as we move to lower levels (e.g. from level 4 to level 5) the number of classes approximately increases by magnitude of ten. Of course, one may wonder how alternative settings would perform. The parameters were selected so there is a balance between the used levels and choosing a different ratio would probably have a negative impact on the performance of the algorithm. For example low influence (e.g. the influence factor was set to 10 instead of 20 that is actually used) would probably not provide enough evidence from the ancestor level to be used in the re-ranking process.

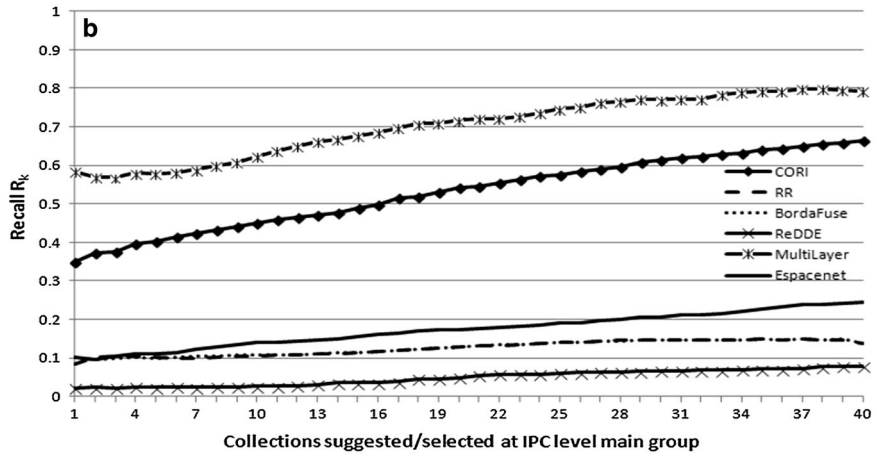
5.2 IPC selection results

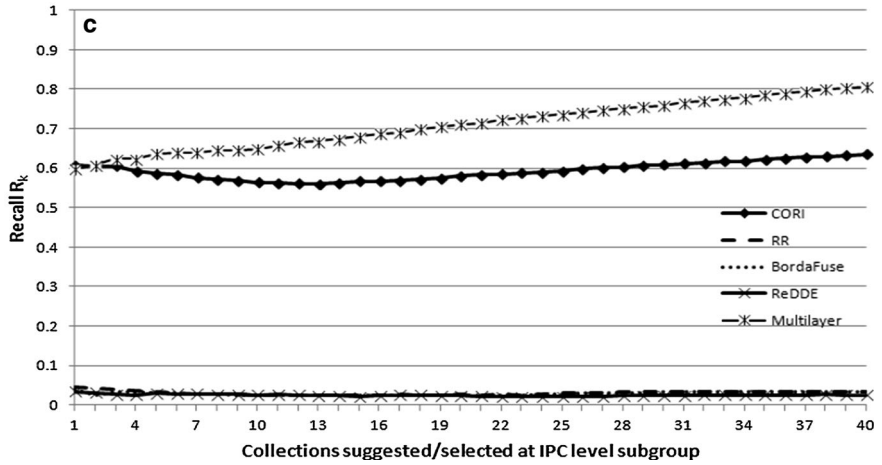
The collection selection algorithms are compared in their ability to automatically suggest IPCs given a query by calculating the R_k. Figure 1 shows the results produced from the source selection algorithms as they gradually select more sub-collections (X axis) at split3 (Fig. 1a), split4 (Fig. 1b) and split5 (Fig. 1c). The results produced from the multilayer algorithm are shown only in Fig. 1b, c because the algorithm was tested only on split4 and split5. The parameter used for the multilayer method was set at 0.2, the value on which the performance of the method was optimized in terms of R_k.

Figure 1a, b also show the IPCs retrieved using the classification search provided by the online Espacenet service.[9] We used the same set of queries and the Espacenet/OPS service

[9] http://worldwide.espacenet.com/classification?locale=en_EP.


a
Recall R_k
CORI
RR
BordaFuse
ReDDE
Espacenet
Collections suggested/selected at IPC level subclass


b
Recall R_k
CORI
RR
BordaFuse
ReDDE
MultiLayer
Espacenet
Collections suggested/selected at IPC level main group


c
Recall R_k
CORI
RR
BordaFuse
ReDDE
Multilayer
Collections suggested/selected at IPC level subgroup

◀**Fig. 1** **a** Results of the source suggestion algorithm(s) at level of subclass. **b** Results of the source suggestion algorithm(s) at level of main group. **c** Results of the source suggestion algorithm(s) at level of subgroup

to retrieve the results illustrated in these figures. The results are significantly worse than the ones produced by CORI and multilayer, however it must be clearly said that the results are not directly comparable with the other IPC suggestion methods. This is because Espacenet/OPS set a limit of 10 terms for their classification service. For this reason we had to use only the titles of the patents/topics while in the DIR runs we used a larger part of the patent/topic to produce the queries that were finally used. It is true that making this necessary compromise makes the results between Espacenet classification search and the rest of the methods presented in this paper not directly comparable. However we included this real system as a useful indication that much better classification search/IPC suggestion services can become available from those which are available in Espacenet and probably in other highly used public service.

The best performing algorithm at split3 (Fig. 1a) is CORI, which identifies more than 95 % of the relevant documents in the first 40 suggested collections, while the other DIR methods identify about 40 %. The Espacenet search service at this split identifies about 55 % of relevant documents in the first 40 suggested collections. At split4 and split5 (Fig. 1b, c) the best performing algorithm is the multilayer where the first 40 suggested collections contain about 80 % of all relevant documents while CORI managed to identify less than 70 %. 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 fusion-based methods ReciRank and BordaFuse produce poor results since they manage to identify only 13 and 3 % of relevant documents in the first 40 selected sub-collections at split4 and split5 respectively.

5.3 Retrieval results

As we discussed earlier in the paper the multilayer method has two main objectives. One objective is to suggest IPCs given a query and help a patent professional to efficiently get a good overview of the technical concepts of a patent. This objective relates more to patent classification task or to patent search task where starting points are needed because the patent searcher is not familiar with the technical area under examination. Another objective is to make a filtered search using the automatically selected IPCs produce a single list of results.

To evaluate the second objective we run the actual patent document retrieval task. Table 5 shows the results of the runs performed on the centralized index and for each DIR method that was tested on split3, split4 and split5. Each DIR run consists of two parts. The first part refers to the method used for the resource selection whereas the second to the results merging method. For example, the *CORI–SSL* run refers to the run for which we used CORI for the resource selection and SSL for the result merging. The DIR methods have been presented in Sect. 2.2. The multilayer algorithm was performed and evaluated on split4 and split5. The performance of the DIR methods should be primarily compared to the centralized run and between them. The optimal method is a retrospective method, which cannot be practically implemented, but it is very useful as it indicates the upper limit for DIR methods and shows the prospects, at least in theory, of these methods.

Table 5 Results of runs at split3, split4 and split5

	10 collections selected			20 collections selected		
	Pres@100	MAP@100	Recall@100	Pres@100	MAP@100	Recall@100
split3						
Optimal–SSL	0.301[a]	0.118[a]	0.386[a]	0.301[a]	0.120[a]	0.386[a]
Centralized	0.257[b]	0.105[b]	0.339[b]	0.257[b]	0.105[b]	0.339[b]
CORI–CORI	**0.261**[b]	**0.106**[b]	**0.342**[b,c]	0.256[b]	0.102[b]	0.336[b]
CORI–SSL	**0.266**[b]	**0.110**[b]	**0.349**[c]	**0.263**[b]	**0.109**[b]	**0.343**[b]
Bordafuse–SSL	0.139[c]	0.062[c]	0.176[d]	0.152[c]	0.070[c]	0.191[c]
ReciRank–SSL	0.137[c]	0.062[c]	0.176[d]	0.151[c]	0.067[c]	0.194[c]
ReDDE–SSL	0.052[d]	0.025[d]	0.062[e]	0.095[d]	0.038[d]	0.095[d]
split4						
Optimal–SSL	0.313[a]	0.128[a]	0.418[a]	0.313[a]	0.128[a]	0.418[a]
Centralized	0.257[b]	0.105[b]	0.339[b]	0.257[b]	0.105[b]	0.339[b]
CORI–CORI	0.203[c]	0.081[c]	0.278[c]	0.213[c]	0.086[c]	0.283[c]
CORI–SSL	0.221[d]	0.091[c]	0.281[c]	0.231[d]	0.097[d]	0.289[c]
Bordafuse–SSL	0.077[e]	0.035[d]	0.098[d]	0.087[e]	0.039[e]	0.108[d]
ReciRank–SSL	0.076[e]	0.037[d]	0.099[d]	0.088[e]	0.039[e]	0.116[d]
ReDDE–SSL	0.024[f]	0.011[e]	0.030[e]	0.042[f]	0.018[f]	0.051[e]
Multilayer–SSL	0.256[b]	0.105[b]	0.338[b]	**0.261**[b]	0.105[b]	**0.341**[b]
split5						
Optimal–SSL	0.346[a]	0.146[a]	0.454[a]	0.351[a]	0.148[a]	0.463[a]
Centralized	0.257[b]	0.105[b]	0.339[b]	0.257[b]	0.105[b]	0.339[b]
CORI–CORI	**0.267**[c]	**0.107**[c]	**0.357**[c]	**0.259**[b]	0.105[b]	**0.347**[b,c]
CORI–SSL	**0.270**[c]	**0.110**[c]	**0.362**[c]	**0.263**[b,d]	**0.107**[b]	**0.349**[b,c]
Bordafuse–SSL	0.030[d,e]	0.020[d]	0.036[d]	0.040[c]	0.028[c]	0.046[d]
ReciRank–SSL	0.035[d]	0.020[d]	0.042[d]	0.044[c]	0.024[c]	0.051[d]
ReDDE–SSL	0.021[e]	0.010[d]	0.028[e]	0.039[c]	0.015[c]	0.047[d]
Multilayer–SSL	**0.269**[c]	**0.106**[c]	**0.364**[c]	**0.267**[d]	0.102[b]	**0.352**[c]

a, b, c, d, e, f: mean scores in the same column and for the same split and for the same collection setting (10/20 collections) without a superscript in common are significantly different. For example, centralized performs statistically better than ReDDE at split3 when 10 collections are selected. However, for the same settings centralized is not statistically better than CORI–CORI

Table 5 shows the performance of the tested methods in terms of PRES, MAP and Recall measures. Measurements in bold report better performance compared to the centralized. The superscripts show the statistical significance of the methods as follows: scores in the same split, same collection size and for the same evaluation metric without a superscript in common are significant different. Significance is tested using a paired *t* test. The optimal approach is not directly compared to the centralized as it is practically impossible to be implemented.

The results presented in Table 5 show that the best performing collection/IPC selection algorithm at the level of subclass (split3) is CORI. CORI performs statistically better than the other DIR methods, however its improvement over the centralized is not statistically

significant. The superiority of CORI in the patent domain as collection selection method compared to BordaFuse and Reciprocal Rank is something consistent with our previous study (Salampasis et al. 2012). However, we observe that CORI does not perform in a similar way as the number of sub-collections/IPCs increases. Specifically, CORI performance deteriorates at split4 while it improves at split5. We believe that this result, which is unexpected at first sight, can be elucidated if we carefully consider the nature and the organisation of the IPC hierarchy. Groups in IPC at level 3 (subclass level) are very homogeneous and patent documents in different technical domains are very well separated between different subclasses (sub-collections) therefore it is very likely that all relevant patents reside in few sub-collections when using IPC at level 3. On the other hand, IPC at levels 4 and below refer to many and very similar technical ideas, therefore are becoming difficult to distinguish between different main groups and subgroups. In fact, the most practical use of IPCs in organisations working with patents occurs at levels 4 and 5; therefore it is not a coincidence that classification search systems have as their "default" level the level of main group or subgroup.

While the previous argument explains the high performance CORI attains at level 3, we believe that better performance of CORI at level 5 in comparison to level 4 should be attributed to the greater effect from the clustering hypothesis. Patents are better distributed to different technical domains at level 5 and, at least in theory, that is the backdrop for any effective resource selection algorithm to perform better. CORI using the hyper-document approach managed to perform well at the level of subgroup regardless the large number of IPC codes because the collections were better distinguished from each other at this level compared to the level of main group.

Further to the previous explanations it should be said that generally resource selection is plagued by uncertainty (Markov et al. 2013) as it is usually based on limited information compared to centralized retrieval. This is also seen in the performance that CORI has at different IPC levels. In fact, our proposed method taking multiple estimates from various IPC levels performs more stable than CORI and to some extent alleviates the "uncertainty" issue that CORI obtains in the results presented in Table 5.

We also notice that the performance of the source selection algorithms that utilize the centralized index (ReDDE, BordaFuse and ReciRank) follow a decreasing trend as the number of IPCs increases. This is in contrast to the performance of the optimal method, which consistently improves as the number of sub-collections increases.

It seems that the best runs at split3 and split5 are those selecting fewer IPCs (10 instead of 20) and requesting more documents (100 instead of 50) from each selected IPC/sub-collection. This fact is probably a combination of the fact that: (a) a small number of relevant documents exists for each topic which are located in a small number of IPCs and (b) resource selection performs well from the beginning (i.e. when selecting few collections) and remains pretty stable, as more collections are selected, at levels 3 and 5. Fact (a) is presented in Table 3, which shows how relevant documents are allocated to IPCs/-sub-collections in each split. Fact (b) is illustrated in Fig. 1a, c.

However, the differences observed are small and this behaviour is not consistent. For example the opposite happens at split4 (selecting 20 IPCs performs better than selecting 10). This can be explained because we observe that the improvement of R_k of CORI and Multilayer when selecting more collections at split4 is substantially larger than at split3 and split5. For example, the improvement for CORI in terms of its performance in selecting 10 and 20 collections measured by R_k is 6.5 % at split3, 16.7 % at split4 and 2.9 % at split5. Having such an improvement at split4 in terms of selecting relevant

collections explains why the retrieval results at split4 are better when selecting 20 IPCs/sub-collections instead of 10.

The most interesting and important finding for this study is that the multilayer method performs better or similar to the other tested methods at lower levels. The multilayer method managed to select more relevant collections than CORI at split4 and split5 by utilising information from higher levels. The performance of actual runs using our source selection method at split4 and split5 is better than using CORI as source selection but also is better or similar compared to the centralized index approach. At split5 multilayer performs significantly better than the centralized method in all tested combinations and in some settings performs also better than CORI—that is when CORI selects 20 IPC subgroups during the resource selection.

Additionally, it is very interesting that some DIR approaches managed to perform better than the centralized approach which is again consistent from our previous studies (Salampasis et al. 2012; Giachanou et al. 2013). 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. However, we must point out that again when resource selection does not begin effectively from the beginning (as it is the case in split4 and we discussed in previous paragraphs in this Section), retrieval results are not better than the centralized run. Apparently when selecting more IPCs (main groups at level 4), the retrieval results would become comparable or even better to the centralized one, however we wanted only to focus on selecting a small number of IPC codes, something that resembles a realistic scenario.

The collections in our study are heterogeneous in size, as shown in Table 2, and therefore ReDDE would be expected to outperform CORI as it is generally observed in DIR research. However, this is not validated in our experiments. We believe that the main reason is that our environment is different from a typical uncooperative environment where ReDDE is indeed superior when compared to CORI in collections with high size variability. Our experiments assume a cooperative environment and therefore we did not apply any source sampling. ReDDE has been proposed and tested for uncooperative environments. Using the centralized index instead of a sample centralized index forced ReDDE to constantly rank the few collections that contain a large number of documents high in the ranking. Potentially, in the future, we can apply source sampling to build a sample centralized index and re-examine the performance of ReDDE and the fusion-based methods.

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 (Willett 1988; Fuhr et al. 2012) and could improve efficiency and effectiveness. Of course in case of patent search, efficiency is not an issue because complete patent collections can be acquired and indexed centrally. However, we believe that searching and browsing on sub-collections rather than the complete collection could potentially 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 (Van Rijsbergen 1979) which states that related documents residing in the same cluster (sub-collection) tend to satisfy same information needs. The clustering hypothesis is proved by Fuhr et al. (2012) 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.

Finally we would like also to mention that a web-based tool implementing classification search using the multilayer method was evaluated in a user study conducted in a national

patent office and which is reported elsewhere in the literature (Giachanou et al. 2014). This study aimed to examine the effectiveness of the multilayer algorithm for classification search, and also if patent examiners can handle the complexity of using the tool for classification search at different levels within a single information seeking episode. From an effectiveness perspective, the results suggest that the patent examiners managed to identify more effectively IPC classification codes using the tool that is based on the multilayer method rather than using the Espacenet classification search service. However these results cannot be considered as conclusive due to the relatively small number of users that participated in the user study (12 patent examiners). Nevertheless the opinions, which are obtained using post-experiment questionnaires, were very positive about (a) generally using the multilayer tool and (b) being able to search at different levels.

6 Conclusion and future work

In this paper, a new multilayer collection selection algorithm was presented with the aim to assist patent examiners in identifying relevant IPC codes and facilitate patent searches. The multilayer method exploits the hierarchical structure of the IPC scheme to automatically suggest IPC codes by utilizing both the topical relevance of IPCs at a particular level of interest and the topical relevance of their ancestors in the IPC hierarchy. The new method was tested on CLEF-IP collection, which was first divided into topically organized sub-collections using the IPC levels of subclass (split3), main group (split4) and subgroup (split5). The new method was compared with state-of-the-art algorithms from DIR domain, fusion-based methods and the centralized approach.

The results showed that the multilayer method performed better than the other tested DIR collection selection algorithms in recall-oriented settings, i.e. multilayer had the best performance in identifying relevant IPCs given a patent query. In relation to the actual patent document retrieval, multilayer performed similar or better not only compared to the rest of the tested methods from DIR domain but more importantly to the centralized approach. In addition to the multilayer, another collection selection method, CORI, managed also to outperform the centralized. On the other side both the fusion-based methods Reciprocal Rank and BordaFuse and ReDDE 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 in comparison to clustering method. We also plan to explore how features such as sub-collections sizes could influence the performance of the multilayer collection selection method. We will further examine the possibility of extending the method to rank the collections exploiting the hierarchical scheme at all available levels.

In conclusion, we feel that the discussion and the experiment presented in this paper are useful for the development of search tools based on DIR methods. We also feel that tools based on the multilayer method can be developed, and these tools will be useful for users of patent search systems who need to utilize the most appropriate search tools given a specific task at hand (e.g. classification search, understanding the technical area of patent, prior-art search). 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 as tools in patent search systems.

Acknowledgments The second author is supported by a Marie Curie fellowship and the research leading to some of the results presented in this paper has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under Grant agreement No. 275522 (PerFedPat).

References

Adams, S. (2000). Using the international patent classification in an online environment. *World Patent Information, 22*(4), 291–300.

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.

Allan, J., Connell, M. E., Croft, B.W., Feng, F.-F., Fisher, D., & Li, X. (2000). Inquery and TREC-9. In *Proceedings of the 9th text retrieval conference* (*TREC'09*), pp. 551–562.

Arguello, J., Diaz, F., Callan, J., & Crespo, J.-F. (2009). Sources of evidence for vertical selection. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval*. MA, USA, pp. 315–322.

Aslam, J. A., & Montague, M. (2001) Models for metasearch. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*. SIGIR'01. New York: ACM, pp. 276–284.

Atkinson, K. H. (2008). Toward a more rational patent search paradigm. In *Proceedings of the 1st ACM workshop on patent information retrieval*. PaIR'08. New York, NY: ACM, pp. 37–40.

Beckers, T., Dungs, S., Fuhr, N., Jordan, M., & Kriewel, S. (2012). ezDL: An interactive search and evaluation system. In *SIGIR 2012 workshop on open source information retrieval*, pp. 9–16.

Beney, J. (2010). LCI-INSA linguistic experiment for CLEF-IP classification track. In *CLEF* (*Notebook Papers/LABs/Workshops*). Padua, Italy.

Benzineb, K., & Guyot, J. (2011) Automated patent classification. In *Current challenges in patent information retrieval*. Berlin: Springer.

Bonino, D., Ciaramella, A., & Corno, F. (2010). Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. *World Patent Information, 32*(1), 30–38.

Buckland, M., & Plaunt, C. (1997) Selecting libraries, selecting documents, selecting data. In *Proceedings of the international symposium on research, development & practice in digital libraries*, pp. 85–91.

Cai, L., & Hofmann, T. (2004). Hierarchical document categorization with support vector machines. In *Proceedings of the thirteenth ACM international conference on information and knowledge management*. CIKM'04. New York: ACM, pp. 78–87.

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

Callan, J., 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). New York, NY: ACM.

Cetintas, S., & Si, L. (2012). Effective query generation and postprocessing strategies for prior art patent search. *Journal of the American Society for Information Science and Technology, 63*(3), 512–527.

Chakrabarti, S., Dom, B., & Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. In *Proceedings of the 1998 ACM SIGMOD international conference on management of data*. SIGMOD'98. New York: ACM, pp. 307–318.

Chen, Y.-L., & Chang, Y.-C. (2012). A three-phase method for patent classification. *Information Processing and Management, 48*(6), 1017–1030.

Chen, Y.-L., & Chiu, Y.-T. (2011). An IPC-based vector space model for patent retrieval. *Information Processing and Management, 47*(3), 309–322.

D'hondt, E. K. L. (2014). *Cracking the patent: Using phrasal representations to aid patent classification*. Dissertation, Radboud Universiteit Nijmegen, Nijmegen, Netherlands.

D'hondt, E., Verberne, S., Koster, C. H. A., & Boves, L. (2013). Text representations for patent classification. *Computational Linguistics, 39*(3), 755–775.

Demeester, T., Trieschnigg, D., Nguyen, D., & Hiemstra, D. (2013). Overview of the TREC 2013 federated web search track. In *TREC*.

Derieux, F., Bobeica, M., Pois, D., & Raysz, J.-P. (2010). Combining semantics and statistics for patent classification. In M. Braschler, D. Harman, & E. Pianta (Eds.). *CLEF* (*Notebook Papers/LABs/Workshops*).

Dirnberger, D. (2011). A guide to efficient keyword, sequence and classification search strategies for biopharmaceutical drug-centric patent landscape searches—A human recombinant insulin patent landscape case study. *World Patent Information, 33*(2), 128–143.

Fall, C. J., Törcsvári, A., Benzineb, K., Karetka, G., & Torcsvari, A. (2003). Automated categorization in the international patent classification. *SIGIR Forum, 37*(1), 10–25.

Fix, E., & Hodges, J. (1951). *Discriminatory analysis. Nonparametric discrimination: Consistency properties*. Randolph Field, Texas: USAF School of Aviation Medicine.

French, J. C., & Powell, A. L. (2000). Metrics for evaluating database selection techniques. *World Wide Web, 3*(3), 153–163.

French, J. C., Powell, A. L., Callan, J., Viles, C. L., Emmit, T., Prey, K. J., & Mon, Y. (1999). 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*). ACM Press, pp. 238–245.

Fuhr, N. (1999). A decision-theoretic approach to database selection in networked IR. *ACM Transactions on Information Systems, 17*(3), 229–249.

Fuhr, N., Lechtenfeld, M., Stein, B., & Gollub, T. (2012). The optimum clustering framework: Implementing the cluster hypothesis. *Information Retrieval, 15*(2), 93–115.

Gey, F., Buckland, M., Chen, A., & Larson, R. (2001). Entry vocabulary—A technology to enhance digital search. In *Proccedings of the 1st international conference on human language technology*, pp. 91–95.

Giachanou, A., Salampasis, M., & Paltoglou, G. (2013). Multilayer collection selection and search of topically organized patents. In *Integrating IR technologies for professional search*.

Giachanou, A., Salampasis, M., Satratzemi, M., & Samaras, N. (2014). A user-centered evaluation of a web based patent classification tool. In *Proceedings of the workshop "beyond single-shot text queries: bridging the gap(s) between research communities" co-located with iConference 2014*.

Guyot, J., Benzineb, K., & Falquet, G. (2010). myClass: A mature tool for patent classification. In M. Braschler, D. Harman, & E. Pianta (Eds.), *Proceedings conference on multilingual and multimodal information access evaluation*. Italy: Padua.

Harris, C., Arens, R., & Srinivasan, P. (2011). Using classification code hierarchies for patent prior art searches. In M. Lupu et al. (eds). *Current challenges in patent information retrieval*. The information retrieval series. Berlin: Springer, pp. 287–304.

Itoh, H. (2005). NTCIR-5 patent retrieval experiments at RICOH. In *Proceedings of NTCIR-5 workshop meeting*. Tokyo.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., & Saarela, A. (2000). Self organization of a massive document collection. *IEEE Transactions on Neural Networks, 11*(3), 574–585.

Kong, Z., & Lalmas, M. (2007). Combining multiple sources of evidence in XML multimedia documents: An inference network incorporating element language models. In *Proceedings of the 29th European conference on IR research*, pp. 716–719.

Konishi, K. (2005). Query terms extraction from patent document for invalidity search. In *Proceedings of NTCIR-5 workshop meeting*.

Kosmopoulos, A., Gaussier, E., Paliouras, G., & Aseervatham, S. (2010). The ECIR 2010 large scale hierarchical classification workshop. *ACM SIGIR Forum, 44*(1), 23–52.

Koster, C., Seutter, M., & Beney, J. (2001). Classifying patent applications with winnow. *Proceedings of Benelearn 2001 Conference* (pp. 19–26). Belgium: Antwerpen.

Krier, M., & Zaccà, F. (2002). Automatic categorisation applications at the European patent office. *World Patent Information, 24*(3), 187–196.

Larkey, L. S. (1998). Some issues in the automatic classification of US patents. In *Working notes for the workshop on learning for text categorization*. Madison, Wisconsin.

Larkey, L. S. (1999). A patent search and classification system. *Proceedings of the fourth ACM conference on digital libraries* (pp. 179–187). New York, NY: ACM.

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*. CIKM'00. McLean, Virginia, USA: ACM New York, NY, USA, pp. 282–289.

Larson, R. R. (2003). Distributed IR for digital libraries. In T. Koch & I. Sølvberg (Eds.) *Research and advanced technology for digital libraries*, 2769, pp. 487–498.

Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research, 5*, 361–397.

Loh, H. T., He, C., & Shen, L. (2006). Automatic classification of patent documents for TRIZ users. *World Patent Information, 28*(1), 6–13.

Lupu, M. (2011). The status of retrieval evaluation in the patent domain. In *Proceedings of the 4th workshop on patent information retrieval—PaIR'11*, p. 31.

Lupu, M., & Hanbury, A. (2013). Patent Retrieval. *Foundations and Trends in Information Retrieval, 7*(1), 1–97.

Magdy, W., & Jones, G. (2010). PRES: A score metric for evaluating recall-oriented information retrieval applications. In *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval*. SIGIR'10. New York, NY: ACM, pp. 611–618.

Markov, I., Azzopardi, L., & Crestani, F. (2013). Reducing the uncertainty in resource selection. *Advances in Information Retrieval*, pp. 507–519.

Moffat, A., & Zobel, J., 1994. Information retrieval systems for large document collections. In *Proceedings of the third text retrieval conference* (*TREC-3*), pp. 85–94.

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 information retrieval*. SIGIR'03, pp. 290–297.

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 and Management, 44*(4), 1580–1599.

Paltoglou, G., Salampasis, M., & Satratzemi, M. (2009). Simple adaptations of data fusion algorithms for source selection. In M. Boughanem et al. (Eds.) *Proceedings of the 31th European conference on IR research on advances in information retrieval*. Lecture Notes in Computer Science. Toulouse, France: Springer, pp. 497–508.

Paltoglou, G., Salampasis, M., & Satratzemi, M. (2011). Modeling information sources as integrals for effective and efficient source selection. *Information Processing and Management, 47*(1), 18–36.

Piroi, F., Lupu, M., & Hanbury, A. (2010). CLEF-IP 2010: Retrieval experiments in the intellectual property domain. In *Notebook papers/LABs/workshops*. Padua, Italy.

Piroi, F., Lupu, M., Hanbury, A., Magdy, W., Sexton, A. P., & Filippov, I. (2012). CLEF-IP 2012: Retrieval experiments in the intellectual property domain. In *CLEF* (*Online working notes/labs/workshop*). Rome, Italy.

Piroi, F., Lupu, M., Hanbury, A., & Zenz, V. (2011). CLEF-IP 2011: Retrieval in the intellectual property domain. In *Cross-language evaluation forum* (*notebook papers/labs/workshop*). Amsterdam, The Netherlands.

Piroi, F., & Zenz, V. (2011) Evaluating information retrieval in the intellectual property domain: The Clef–Ip campaign. In M. Lupu et al. (Eds.) *Current challenges in patent information retrieval*. The Information retrieval series. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 87–108.

Porter, M. (1980). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems, 14*(3), 130–137.

Powell, A. L., & French, J. C. (2003). Comparing the performance of collection selection algorithms. *ACM Transactions on Information Systems, 21*(4), 412–456.

Powell, A. L., French, J. C., Callan, J., Connell, M., & Viles, C. L. (2000). The impact of database selection on distributed searching. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*, pp. 232–239.

Roda, G., Tait, J., Piroi, F., & Zenz, V. (2009). CLEF-IP 2009: Retrieval experiments in the intellectual property domain. In C. Peters et al. (Eds.) *CLEF working notes 2009*, 6241, pp. 385–409.

Salampasis, M., Paltoglou, G., & Giahanou, A. (2012). Report on the CLEF-IP 2012 experiments: Search of topically organized patents. In *Proceedings of CLEF Conference*.

Shokouhi, M. (2007). Central-rank-based collection selection in uncooperative distributed information retrieval. In *Proceedings of the 29th European conference on information retrieval*. ECIR'07. Rome: Springer, pp. 160–172.

Shokouhi, M., & Si, L. (2011). Federated search. *Foundations and Trends in Information Retrieval, 5*(1), 1–102.

Si, L., & Callan, J. (2003a). A semisupervised learning method to merge search engine results. *ACM Transactions on Information Systems, 21*(4), 457–491.

Si, L., & 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*. Toronto, Canada: ACM New York, NY, USA, pp. 298–305.

Si, L., & Callan, J. (2005). Modeling search engine effectiveness for federated search. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*. Salvador: ACM New York, NY, USA, pp. 83–90.

Si, L., Jin, R., Callan, J., & Ogilvie, P. (2002). A language modeling framework for resource selection and results merging. In *Proceedings of the eleventh international conference on information and knowledge management*. ACM Press, pp. 391–397.

Sigurbjörnsson, B., Kamps, J., & de Rijke, M. (2004). Multiple sources of evidence for XML retrieval. *Proceedings of the 27th annual international conference on research and development in information retrieval—SIGIR'04* (pp. 554–555). New York: ACM Press.

Tikk, D., Biró, G., & Törcsvári, A. (2007). A hierarchical online classifier for patent categorization. In H. A. do Prado & E. Ferneda (Eds.), *Emerging technologies of text mining*. IGI Global: Hershey, PA.

Van Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworth-Heinemann.

Verberne, S., & D'hondt, E. (2011). Patent classification experiments with the linguistic classification system LCS in CLEF-IP 2011. *CLEF* (*notebook papers/labs/….*

Vijvers, W. G. W. (1990). The international patent classification as a search tool. *World Patent Information, 12*(1), 26–30.

Voorhees, E. M., Gupta, N. K., & Johnson-Laird, B. (1994). The collection fusion problem. In D. K. Harman (Ed.) *Proceedings of the 3rd text retrieval conference TREC3*. National Institute of Standards and Technology, pp. 95–104.

Willett, P. (1988). Recent trends in hierarchic document clustering: a critical review. *Information Processing and Management, 24*(5), 577–597.

Wolter, B. (2012). It takes all kinds to make a world—Some thoughts on the use of classification in patent searching. *World Patent Information, 34*(1), 8–18.