Knowledge and Information Systems (2019) 61:631–660
https://doi.org/10.1007/s10115-018-1322-7

SURVEY PAPER

Patent retrieval: a literature review

Walid Shalaby[1] · **Wlodek Zadrozny**[1]

Received: 13 November 2017 / Revised: 10 December 2018 / Accepted: 17 December 2018 /
Published online: 14 January 2019

Abstract
With the ever increasing number of filed patent applications every year, the need for effective and efficient systems for managing such tremendous amounts of data becomes inevitably important. Patent retrieval (PR) is considered the pillar of almost all patent analysis tasks. PR is a subfield of information retrieval (IR) which is concerned with developing techniques and methods that effectively and efficiently retrieve relevant patent documents in response to a given search request. In this paper, we present a comprehensive review on PR methods and approaches. It is clear that recent successes and maturity in IR applications such as Web search cannot be transferred directly to PR without deliberate domain adaptation and customization. Furthermore, state-of-the-art performance in automatic PR is still around average in terms of recall. These observations motivate the need for interactive search tools which provide cognitive assistance to patent professionals with minimal effort. These tools must also be developed in hand with patent professionals considering their practices and expectations. We additionally touch on related tasks to PR such as patent valuation, litigation, licensing, and highlight potential opportunities and open directions for computational scientists in these domains.

Keywords Information retrieval · Patent retrieval · Patent mining · Patent prior art search · Survey

1 Introduction

Patents represent proxies for economic, technological, and even social activities. The Intellectual Property (IP) system motivates the disclosure of novel technologies and ideas by granting inventors exclusive monopoly rights on the economic value of their inventions. Patents, therefore, have a major impact on enterprises market value [72]. With the continuous rise in the number of filed patent applications every year, the need for effective and efficient systems for managing such tremendous amounts of data becomes inevitably important.

✉ Walid Shalaby
wshalaby@uncc.edu

Wlodek Zadrozny
wzadrozn@uncc.edu

[1] Department of Computer Science, University of North Carolina at Charlotte, Charlotte, USA

Typical patent analysis tasks include: (1) technology exploration in order to capture new and trendy technologies in a specific domain, and subsequently using them to create new innovative services, (2) technology landscape analysis in order to assess the density of patent filings of specific technology, and subsequently direct R&D activities accordingly, (3) competitive analysis and benchmarking in order to identify strengths and differences of corporate's own patent portfolio compared to other key players working on related technologies, (4) patent ranking and scoring in order to quantify the strength of the claims of an existing or a new patent, and (5) prior art search in order to retrieve patent documents and other scientific publications relevant to a new patent application. All those patent-related activities require tremendous level of domain expertise which, even if available, must be integrated with highly sophisticated and intelligent analytics that provide cognitive and interactive assistance to the users.

Patent retrieval (PR) is the pillar of almost all patent analysis tasks. PR is a subfield of Information Retrieval (IR) which is concerned with developing techniques and methods that effectively and efficiently retrieve relevant patent documents in response to a given search request. Although the field of IR has received huge advances from decades of research and development, research in PR is relatively newer and more challenging. On the one hand, patents are multi-page, multi-modal, multi-language, semi-structured, and metadata rich documents. On the other hand, patent queries can be a complete multi-page patent application. These unique features make traditional IR methods used for Web or ad hoc search inappropriate or at least of limited applicability in PR.

Moreover, patent data are multi-modal and heterogeneous. As indicated by Lupu et al. [52], analyzing such data is a challenging task for many reasons; patent documents are lengthy with highly complex and domain-specific terminology. To establish their work novelty, inventors tend to use jargon and complex vocabulary to refer to the same concepts. They also use vague and abstract terms in order to broaden the scope of their patent protection making the problem of patent analysis linguistically challenging.

PR starts with a search request (query) which often represents a patent application under novelty examination. Therefore, several methods for query reformulation (QRE) have been proposed in order to select, remove, or expand terms in the original query for improved retrieval. QRE methods are keyword-based, semantic-based, or interactive. Keyword-based methods work by searching for exact matches between search query terms and the target corpus, and thus fail to retrieve relevant documents which use different vocabularies but have similar meaning to the reformulated query. In order to alleviate the vocabulary mismatch problem, semantic-based methods try to search by meaning through expanding queries and/or target corpus with similar or related terms and thus bridging the vocabulary gap. Because neither method proved acceptable performance, few interactive methods were proposed to allow users to interactively control QRE with reasonable effort.

This review aims to provide researchers with an illustrative and critical overview of recent trends, challenges, and opportunities in PR. The rest of this paper is organized as follows. Section 2 presents some preliminaries and background about patents data. Section 3 provides an overview of evaluation tracks and data collections for PR benchmarking. An illustration of PR tasks is presented in Sect. 4. Section 5 presents a comprehensive review on PR methods and approaches. Section 6 lightly touches on related tasks such as patent quality assessment, litigation, and licensing. Finally, concluding remarks are presented in Sect. 7.

Table 1 Patent kind codes of major patent offices

Type	USPTO (US)	EPO (EP)	WIPO (WO)
A1	application	application w/ search report	
A2	republished application	application w/o search report	
A3	-	search report	
A4	-	supplementary search report	publication of amended claims
A9	modified application		
B1	granted patent w/o application	granted patent (publication)	-
B2	granted patent w/ application	amended B1	-

2 Preliminaries

2.1 Patent documents and kind codes

Patent documents are mostly textual. They are highly structured with typical elements (sections) including *title, abstract, background of the invention, description and claims*. The *description* section articulate in details the technical specification of the invention and its possible embodiments. The *claims* section is the most significant one as it describes the scope of protection sought by the inventor and hence encodes the real value of the patent. Patent documents are lengthy with highly complex and domain-specific terminology. They also contain multiple data types (e.g., text, images, flow charts, formulae) with a rich set of metadata and bibliographic information (e.g., *classification codes*, *citations*, *inventors*, *assignee*, *filing/publication dates*, *addresses*, *examiners*).

Typically, each patent has a set of pertaining documents which are published throughout its life cycle. All documents are identified by an alphanumeric name with a common naming convention. Names start with two letters identifying the issuing patent office (e.g., US and EP), then the patent number as sequence of digits, and finally a suffix indicating the document's kind code. The kind code identifies the stage in the patent life cycle at which the document is published. Table 1 shows a brief description of kind codes used at major patent offices and organizations including the US Patent and Trademark Office[1] (USPTO), the European Patent Office[2] (EPO) and the World Intellectual Property Organization[3] (WIPO).

2.2 Patent classification

Patent offices organize patents by assigning classification codes to each of them based on the technical features of the invention. The patent classification system is a hierarchical one. Common classification systems include the International Patent Classification (IPC) and the Cooperative Patent Classification (CPC). The CPC was jointly developed by the USPTO and EPO to replace the US Patent Classification (USPC) and European classification system (ECLA).

[1] http://www.uspto.gov/.

[2] http://www.epo.org/.

[3] http://www.wipo.int/portal/en/index.html.

2.3 Patent families

A patent family is a collection of patents that refer to the same invention and are granted at different countries around the world [74]. Typically, they describe a single invention in different languages depending on the issuing patent office. In the context of PR and prior art search, patents belonging to the same patent family could be used to expand the prior art list of the topic patent as they disclose the same invention.

3 Data and evaluation tracks

This section presents an overview of evaluation tracks organized for patent data analysis along with available data collections with focus on tasks pertaining to PR.

3.1 CLEF-IP collections

The Conference and Labs of the Evaluation Forum[4] (CLEF) is a European series of workshops which started in 2001 to foster research in Cross Language Information Retrieval (CLIR). The Intellectual Property (IP) track (CLEF-IP) which ran between (2009 and 2013) was organized to: (1) foster research in patent data analysis and (2) provide large and clean test collections of multi-language patent documents, specifically in the three main European languages (English, French, and German). Research labs have the opportunity to test their methods on multiple shared tasks such as PR, patent classification, image-based PR, image classification, flowchart recognition, and structure recognition [71–74,77].

The CLEF-IP data collection are patent documents extracted from USPTO, EPO and WIPO data. It is provided through the Information Research Facility[5] (IRF) and hosted by Marec.[6] Patent documents are provided in XML format and have common Document Type Definition (DTD) schema. The collection was constructed according to the proposed methodology by Graf and Azzopardi [27] and is divided into two pools:

1. *The corpus pool*: Documents selected from this pool are provided for participating labs as training or lookup instances.
2. *The topics pool*: Documents selected from this pool are called topics, and they represent testing or evaluation instances. For example, in prior art search, the topic might be a patent application document for which it is required to retrieve prior art. In this scenario, the patents that constitute prior art are called relevance assessments and obtained from the corpus pool.

The XML documents consist of the main textual sections such as *bibliographic data*, *abstract*, *description*, and *claims*. Each section is written in one or more languages (English, French, and/or German) and is denoted by a language code. At least the claims of granted patents (B1 documents) are written in the three languages because it is EPO requirement once a patent application is granted.

CLEF-IP 2009 collection: this dataset was designed for the prior art search task [77]. The corpus pool contains documents published between (1985 and 2000) (∼2 m documents pertaining to ∼1 m unique patents). The topics pool contains documents published between (2001 and 2006) (∼0.7 m documents pertaining to ∼0.5 m individual patents). Top-

[4] http://www.clef-initiative.eu/.

[5] http://www.ir-facility.org/.

[6] http://www.ifs.tuwien.ac.at/imp/marec.shtml.

ics are sets of documents from the topics pool with sizes ranging from 500 to 10,000 topics. Topics were assembled from granted patent documents including *abstract*, *description*, and *claims* sections. Citation information from the *bibliographic data* section was excluded.

A major pitfall in this dataset is its topics, which were chosen from granted patent documents (B1 documents). Initially, the creators of the dataset were motivated by having topics from granted patent documents which have claims in three languages. This was thought to provide a kind of parallel corpus suitable for CLIR. The problem of using such documents is simple, it contradicts the practice of IP search professionals who start with the patent application document not the granted one.

CLEF-IP 2010 collection: this dataset was created for the prior art search and patent classification tasks [71]. The corpus pool of this dataset contains documents with publication date before 2002 ($\sim$2.6 m documents pertaining to $\sim$1.9 m unique patents). The topics pool contains documents published between (2002 and 2009) ($\sim$0.8 m documents pertaining to $\sim$0.6 m unique patents). Topics for the prior art task are two sets of documents from the topics pool; a small set of 500 topics and a larger set of 2000 topics. Unlike the CLEF-IP 2009 dataset, topics are assembled from patent application documents rather than granted patent documents.

CLEF-IP 2011 collection: this dataset was created as a test collection for four tasks: prior art search, patent classification, image-based prior art search, and image classification [72]. The topics and corpus pools were the same as in CLEF-IP 2010 dataset. For the prior art task, 3973 topics were provided as a separate archive of patent application documents.

CLEF-IP 2012 collection: this dataset was created as a test collection for three tasks: passage retrieval starting from claims, chemical structure recognition, and flowchart recognition [73]. The topics and corpus pools were the same as in CLEF-IP 2010 dataset. The passage retrieval task is designed differently from previous CLEF-IP prior art search collections. The purpose for these tasks is to retrieve both documents and passages relevant to a set of claims. Topics for the passage retrieval task were extracted from patent applications published after 2001. Relevance judgments were the highly relevant citations only (i.e., marked X or Y) in the examiners' search reports (A4 documents) of chosen topic patents.

CLEF-IP 2013 collection: this dataset was created as a test collection for two tasks: (1) passage retrieval from claims and (2) structure recognition from patent images [74]. The topics and corpus pools were the same as in CLEF-IP 2010 dataset. Similar to CLEF-IP 2012, the CLM task is designed to retrieve both documents and passages relevant to a set of claims. Topics for the passage retrieval task were extracted from patent applications published after 2002. Overall, the topics set contained 148 topics extracted from 69 patent applications.

3.2 NTCIR collections

The Japanese National Institute of Informatics Testbeds and Community for Information access Research project[7] (NTCIR) started in 1997 to support research in IR and other areas, focusing on CLIR. NTCIR has been organizing a series of workshops providing test collections to researchers for evaluating their methodologies on multiple CLIR tasks NTCIR [68]. Between NTCIR-3 and NTCIR-11 (2002–2013), there has been dedicated tasks for patent data analysis including patent retrieval [38], classification, mining, and translation.

NTCIR-3: the PR task in NTCIR-3 targeted the "technology survey" problem. The dataset for this task includes: (1) full text of Japanese patent applications between (1998 and 1999), (2) *abstract* of Japanese patent applications between (1995 and 1999) along with their respective

[7] http://research.nii.ac.jp/ntcir.

English translations, and (3) 30 search topics where each topic includes a related newspaper article. The task is to retrieve patents relevant to news articles. Both cross-genre experiments in which patents were retrieved by a newspaper clip as well as ordinary ad hoc retrieval of patents by topics were conducted [38].

NTCIR-4: two PR tasks were organized in NTCIR-4 [20]: (1) patent map generation and (2) invalidity search. The dataset for the PR tasks includes: (1) unexamined Japanese patent applications published between (1993 and 1997) along with English translations of the *abstract*, and (2) 34 search topics where each topic is a claim of a rejected patent application which was invalidated because of existing prior art. Relevance judgments were individual patents that can invalidate a topic claim by its own or in conjunction with other patents. Relevant passages to the invalidated claim were also annotated and added to the relevance judgments.

NTCIR-5: two PR tasks were organized in NTCIR-5 [21]: (1) document retrieval (invalidity search), and (2) patent passage retrieval. The dataset for the invalidity search task includes: (1) unexamined Japanese patent applications published between (1993 and 2002) along with English translations of the *abstract*, and (2) 1200 search topics where each topic is a claim of an invalidated patent application. Relevance judgments were generated in a manner similar to the one used in NTCIR-4 invalidity search task.

NTCIR-6: two PR tasks were organized in NTCIR-6 [22]: (1) Japanese retrieval (invalidity search), and (2) English retrieval. The dataset for the Japanese retrieval task is the same one used in NTCIR-5, but more topics were used (1685 topics). The English retrieval task was focusing on finding all the citations cited by the applicant and the examiner. The dataset for these tasks includes: (1) granted patents from the USPTO between (1993 and 2000) and (2) 3221 search topics where each topic is a granted patent published between (2000 and 2001).

3.3 TREC-CHEM collections

The TREC-CHEM track was organized to motivate large-scale research on chemical datasets, especially chemical patent retrieval [50].

TREC-CHEM 2009: this collection was created as a test collection for two tasks [50]: (1) technology survey, and (2) prior art search. Eighteen topics were provided for the technology survey task where relevance judgments were obtained from experts and chemistry graduate students. For the prior art search, 1000 patents were provided as test topics where relevance judgments were collected from the citations of topic patents as well as their family members. The search corpus contains $\sim$1.2 m chemical patents filed until 2007 at EPO, USPTO, and WIPO. It also contains 59K scientific articles.

TREC-CHEM 2010: this collection was created for the same two tasks as in TREC-CHEM 2009 [51]. Thirty topics were provided for the technology survey task. The search corpus contains $\sim$1.3 m chemical patents and 177K scientific articles. Relevance judgments were created the same way as in TREC-CHEM 2009.

TREC-CHEM 2011: this collection was created for the same two tasks as in previous TREC-CHEM tracks besides a new chemical image recognition task. The technology survey task topics were biomedical and pharmaceutical patents [53].

3.4 Other sources

Other IP data sources are detailed by Schwartz and Sichelman [81]. These include full patent texts as well as bibliographic information from major patent offices such as the USPTO,

EPO, and WIPO. Bibliographic information for patents published from 1976 to 2006 is provided through the National Bureau of Economic Research[8] (NBER) and subsequently cleaned and extended to include patents until 2013.[9] Patent prosecution histories are available through the Patent Application Information Retrieval[10] (PAIR). Patent assignments, filings, classifications, and petition decisions are also provided through the USPTO bulk downloads previously hosted by Google[11] and now by the USPTO.[12]

4 Patent retrieval tasks

The goal of PR is to retrieve relevant patent documents to a given search request (query). This request can take different forms such as a sequence of keywords, a memo, or a complete text document (e.g., a patent application). The purpose of this task is manifold, for example:

- Retrieve related patents to a given patent application in order to gather related work or invalidate one or more of its claims.
- Explore patent filing activity under specific technology.
- Explore the competitive landscape of a given company by looking at other companies filing patents similar to the given company patents.

Because of these multiple objectives, various PR tasks were proposed to fulfill each objective, and multiple datasets were provided depending on the given task.

Prior art search is the main theme of the CLEF-IP and NTCIR tracks. The importance of this task stems from the requirement by all patent offices that filed patents must constitute novel, non-obvious, and non-abstract ideas. Therefore, an important activity through the patent life cycle is to thoroughly ensure that no earlier published patent or material describing the prescribed ideas exist. The task can be defined as follows:

Problem: given a patent application X, retrieve all related documents to X

Prior art search is a total-recall task,[13] therefore it demonstrates several challenges. Search coverage is one of the main challenges, because it is required to cover all previously published material (patent or non-patent literature) in all forms (electronic or printed) which is infeasible. Another major challenge is the need to search through materials written in different languages. Last but not least, traditional IR methods perform poorly when confronted with the patent prior art search task. Mainly because the patent language is full of jargon and user-defined terminology. Inventors intentionally tend to use different vocabularies to express same or similar ideas in order to establish the novelty of their work.

Prior art search is performed at different stages of the patent life cycle, by different stakeholders, for various purposes, and for limited period of time. Understanding the real-life practices of patent professionals is critical to better satisfy their information need [41]. In other words, the search scenario depends on when it is done, by whom, and for what reason(s). Table 2 shows these various scenarios which are detailed below.

Related work search: during the pre-grant stage, inventors and prosecutors run related work search to retrieve all relevant work to the invention. Moreover, some patent offices

[8] https://sites.google.com/site/patentdataproject/.

[9] http://rosencrantz.berkeley.edu/batchsql/.

[10] http://portal.uspto.gov/pair/PublicPair.

[11] https://www.google.com/googlebooks/uspto-patents.html.

[12] https://www.uspto.gov/learning-and-resources/bulk-data-products.

[13] It is required to achieve 100% recall at acceptable precision.

Table 2 Scenarios of patent prior art search

Search task	Who	When	Purpose	Output
Related work	Inventor/prosecutor	Pre-grant	All related work	Applicant's disclosure
Patentability	Prosecutor/examiner	Pre-grant/examination	Novelty breaking work	Grant/modify/reject
Infringement	Owner/investor	Post-grant	Relevant claims/infringing products	Sue/license/clearance
Freedom to operate	Investor	Post-grant	Relevant claims/related work	Clearance
Invalidity	Competitor/defendant	Post-grant	Novelty breaking work	Reexamine/inter-parts review/post-grant review
Technology survey	Technology analyst	Pre/post-grant	All published patents	Survey report

request from inventors an applicant's disclosure document specifying all related publications when filing a new application.

Patentability search: during the examination stage, patent examiners perform patentability search in order to ensure that the proposed ideas are novel, non-obvious, and non-abstract. The output of this task would be a search report with all retrieved relevant publications. In this report, each entry will have a special code indicating whether it is just a related publication or novelty breaking one. Examiners would also specify which passages or figures in retrieved publications constitute relevancy. Depending on the search findings, the patent office might grant, reject, or ask the applicant to modify the patent application. Patentability search is also performed by patent prosecutors as a sanity check. Although this task should be of equal interest to prosecutors who file the patent application as it is to examiners, prosecutors often do not dig deep searching for relevant publications, and delegate finding relevant prior work to examiners in order to save costs.

Infringement search: this task, also called product clearance search, aims to ensure whether an existing or a proposed product is infringing any published patent claim(s). Patent owners require that type search to find out whether a third party has a product with features that are within the scope of one or more claims of their patents. If so, they might either sue or negotiate a license with that infringing party.

Investors and R&D managers, on the other hand, require that type of search to ensure newly proposed product(s) are not infringing a published patent claim(s) and investment in such products would be lucrative. The scope of search in this case would be limited to patent and the copyrighted literature only. Deep understanding and correct interpretation of patent claims are imperative for building the correct correspondence between product features and claims in order to establish or dismiss infringement.

Freedom to operate search: this PR task extends beyond infringement search. Here, investors and R&D managers not only need to make sure that proposed products do not infringe an existing patent or copyrighted material, but also to ensure they have the freedom to file patents on these products without worrying about previous prior art that might invalidate such inventions. Another objective of freedom to operate search is to make better investment decisions and R&D plans according to existing prior art.

Invalidity search: as patents guarantee monopoly rights to their owners on the economic value of granted inventions, companies and other parties usually monitor granted patents of their competitors or pertaining to their technology landscape to ensure competitive superiority. Therefore, invalidity search is performed to find published material that was missed by the patent office during patentability search. Invalidity search is also considered as the first line of defense when a party is confronted with patent infringement lawsuit. Again published material might include patent or non-patent literature such as books, news articles, academic periodicals. After finding such validity breaking material, a third party might file a post-grant (opposition) procedure depending on the patent office policies. For example, the USPTO provides procedures such as reexamination, inter-partes Review,[14] and post-grant review[15] in front of the Patent Trial and Appeal Board[16] (PTAB).

Technology survey: Another PR task where, in a typical scenario, business managers would request search professionals to prepare a survey of patent documents given a memorandum they prepared from some source (e.g., news article). In the PR task at NTCIR-3 [38], this

[14] http://www.uspto.gov/patents-application-process/appealing-patent-decisions/trials/inter-partes-review.

[15] http://www.uspto.gov/patents-application-process/appealing-patent-decisions/trials/post-grant-review.

[16] https://ptabtrials.uspto.gov.

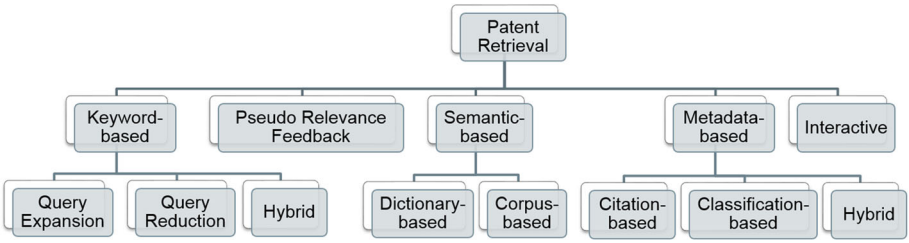

Fig. 1 Taxonomy of patent retrieval methods

basic scenario was limited to the patent literature and it was assumed that patent documents are just a collection of technical papers.

5 Patent retrieval methods

In this section, we present a comprehensive review of PR methods and approaches. We start by presenting available test collections and evaluation metrics. Then, we provide a taxonomy of these approaches highlighting their characteristics and limitations.

As shown in Fig. 1, PR methods can be categorized depending on which piece(s) of data from both the search queries and the search corpus are used for retrieving relevant documents. Keyword-based methods utilize only terms from search queries and look for exact matches in the target corpus. Pseudo-relevance feedback methods utilize terms from the top ranked results of running the initial query to improve the set of relevant retrieved results. Semantic-based methods try to overcome the vocabulary mismatch problem between the search terms and related patents vocabulary by matching them based on their meanings. Metadata-based methods exploit the language-independent non-textual metadata and bibliographic information in order to improve patent retrievability. Finally, interactive methods aim to better organize and present search results to the users. Moreover, through interaction, users are engaged in an iterative process of searching, reviewing, and refining hoping to retrieve as many relevant results as possible.

5.1 Test collections and evaluation measures

As we highlighted in Sect. 3, several datasets were created to support evaluating different PR techniques. In almost all of these datasets, relevant documents to search queries were collected from the citations of topic patent documents (e.g., CLEF-IP 2009/2010/2011 collections). Because these citations represent related prior work, they are appropriate only for the related work search task.

In other datasets such as CLEF-IP 2012/2013 collections, relevant documents were collected from novelty breaking citations found in examiners' search reports, and therefore, these datasets are appropriate for the patentability and invalidity search tasks although invalidity search requires the non-patent literature as well.

Standard information retrieval as well as patent retrieval-specific evaluation measures are generally used to evaluate patent retrieval systems including:

1. Precision (P) and Recall (R) at top-K ranks (e.g., $K = \{1, 5, 10, 50, 100, 1000\}$).

2. Mean Average Precision (MAP) [4] which generally favors early retrieval of relevant documents with less focus on recall.
3. Normalized Discounted Cumulative Gain (nDCG) [39] which favors not only early retrieval of relevant documents but also the respective ranking quality of these documents.
4. Patent Retrieval Evaluation Score (PRES) [56] which was proposed specifically for recall-oriented tasks such as PR. PRES focuses on the overall system recall as well as user's review effort which can be estimated from the rankings at which relevant documents are retrieved.

5.2 Query reformulation (QRE)

The most widely used techniques for patent retrieval are the Query Reformulation (QRE) techniques. These methods aim at transforming the input query Q into $\bar{Q}$ by means of reduction or expansion of Q terms in order to improve the retrievability of relevant documents. QRE can be performed through:

- *Query reduction (QR)*: where a representative subset of terms are selected from Q and used as $\bar{Q}$ terms. Position-based methods are the most commonly used in this category where terms from specific parts or sections of the patent document are used, or given higher matching weight than others. Another example of query reduction is the IPC-based methods which utilize terms from IPC definitions as a lexicon or stop-words list for Q.
- *Query expansion (QE)*: where representative terms other than the ones in Q are extracted and merged with Q to form $\bar{Q}$. Pseudo-relevance feedback (PRF) methods are the most prominent in this category where terms from top ranked results of running Q are used to expand Q terms assuming these top results are relevant [8]. Other semantic-based query expansion methods work by expanding Q with terms of similar meanings such as synonyms or hyponyms.
- *Hybrid (query expansion and reduction)*: where irrelevant terms are removed from Q and more relevant terms are appended to Q to form $\bar{Q}$. Most techniques used for query expansion are appropriate for query reduction as well, where only terms appearing in the expansion list are kept and all others are pruned.

5.2.1 Keyword-based methods

This set of techniques retrieves relevant documents by looking for exact matches between search query term(s) and the target data. keyword search operates under the closed vocabulary assumption where vocabulary is derived solely from terms that appear in the target search data. Table 3 shows some keyword-based methods along with their performance results on benchmark datasets. Keyword-based techniques differ in: (1) which elements of the target data are indexed, (2) which query terms are selected/removed, (3) the relative weights of such terms, and (4) the match scoring function.

Query reduction (QR): the rationale behind QR approaches is intuitive as patents are very long documents with several sections. Querying with the whole document would be impractical and inefficient. Some query reduction methods are position-based; they select relevant terms based on their position in the patent document [15,58,64,94,98]. For example, Verberne and D'hondt [94] used only terms from the *claims* section on the CLEF-IP 2009 collection. However, the results were moderate in terms in MAP compared to other runs on the same collection.

Table 3 Keyword-based patent retrieval methods

Method	Description	Dataset	MAP	P	R	PRES
Verberne and D'hondt [94]	– Remove stop-words and punctuation – Use claims as BOW	clef-ip 2009	0.05	0.01	0.22	–
Magdy et al. [58]	– Remove stop-words and frequent terms – Use different sections with manual weights – Perform IPC filtering – Use bigrams with tf>1	clef-ip 2009	0.12	–	0.63	–
Mahdabi et al. [64]*	– Use query language models on different sections – Use queries of 100 terms – Perform IPC filtering	clef-ip 2010	0.12	–	0.60	0.49
Wang and Lin [99]*	– Use linguistic-based concepts – Concept weighting using weighted tf-idf and mutual information	clef-ip 2010	0.10	–	0.48	0.40
Konishi [43]	– Patterns to identify *claim* components terms – Patterns for explanation terms from *description* – Rank boosting based on IPC	ntcir-5	0.20	–	–	–

*Indicates scores @1000

Magdy et al. [58] experimented using text from different sections of the topic patent on the CLEF-IP 2009 collection. The authors used various combinations of sections including: (1) short sections such as *title, abstract*, first line of the *description*, first sentence of the *claims*, and (2) lengthy sections such as the *description* and the *claims*. The authors assigned different weights to each section manually. Their best scores were achieved using a combination of all short sections and post-filtering retrieved documents keeping only those that share the same IPC classification code with the topic patent. The main challenge with such approach is how to assign the respective weight of each section automatically. Moreover, IPC filtering wouldn't be possible when only partial patent application is available for prior art search.

Mahdabi et al. [64] proposed a position-based query reduction method which selects relevant query terms by building two query language models using various sections of the topic patent: (1) a variant of the weighted log-likelihood model [67], and (2) a model based on the parsimonious language model [36]. Their experiments showed that queries constructed from terms in the *description* section using weighted log-likelihood give better results than other sections which agree with the previous results [6,55,101]. The main advantage of this approach is that respective weights of query terms are derived automatically from the query model. However, some challenges still exist regarding tuning the model parameters such as the smoothing parameter which was set heuristically.

Query expansion (QE): pattern-based QE was proposed in many studies [43,70,99]. Wang and Lin [99] proposed patterns in the form of syntactic rules in order to extract query terms as weighted concepts. Konishi [43] proposed a pattern-based query expansion method for the patent invalidity search task on the NTCIR-5 collection. In this task, the initial query of the topic patent was the terms in the *claims* section. However, rather than using only raw *claims* terms which are often abstract, Konishi [43], using pattern matching, identifies other specific terms in the *description* and uses them as expansion terms. First, components of the invention are extracted from the topic *claim* using handcrafted patterns. Secondly, explanation sentences describing components of the invention are extracted from the *description* using handcrafted patterns. Thirdly, terms from first and second steps are used as the new query. The results showed that this query expansion approach works better than using terms extracted from the *claims* section only. The main drawback of this method is its dependency on manually coded patterns to identify potential terms. Meanwhile, it demonstrates the potential of using entities and their relations as retrieval features motivating the need for deeper and more generic linguistic analysis of patent texts.

5.2.2 Pseudo-relevance feedback (PRF)

These methods are one of the prominent techniques used for QRE. PRF starts with an initial run of the given query Q. Then, terms from top ranked results are used to select, remove, and/or expand terms in Q, assuming that these top results are relevant. PRF is thus advantageous as it works automatically without human intervention but might be computationally inefficient, especially with long queries. Table 4 shows some PRF methods along with their performance results on benchmark datasets.

Despite their effectiveness and popularity, several challenges arise when it comes to PRF-based QRE [6] such as: (1) which part(s) of the patent application should be used as the initial query?; (2) which part(s) of the retrieved results should be used as the source of expansion and/or reduction?; (3) what is the best length of the expansion list in case of query expansion, or the best threshold for removing terms in case of reduction?; (4) which pseudo-relevant results are really relevant and how many of them should be used?; and (5) what is the best

Table 4 Pseudo-relevance feedback patent retrieval methods

Method	Description	Dataset	MAP	PRES
Bouadjenek et al. [6]	– Use different methods of query expansion and reduction from the PRF set – Use Rocchio, MMR, LM	clef-ip 2010 clef-ip 2011	0.13 0.10	0.55 0.45
Magdy et al. [58]	– Naive PRF – Remove stop-words – Use most frequent terms	clef-ip 2009	0.05	–
Mahdabi and Crestani [60]	– Build regression model using relevance score, RF similarities, etc – Use the model to estimate the effectiveness of RF – Use top 100 RF and maximize AP	clef-ip 2010	0.16	0.56
Ganguly et al. [23]	– Perform query segmentation – Retain segments highly to be generated using RF LM	clef-ip 2010	0.14	0.47
Golestan Far et al. [26]	– Manually annotate one relevant RF result – Add terms in the annotated result to the query	clef-ip 2010	0.29[a]	–
Golestan Far et al. [26]	– Assume relevant RF results are known – Add terms more frequent in relevant than irrelevant RF to query	clef-ip 2010	0.48[b]	–

[a] This is a semi-supervised performance.
[b] This is an Oracle performance.

relevance scoring model for the search task (e.g., BM25 [76], the vector space model with tf-idf weighting).

Bouadjenek et al. [6] provided a thorough evaluation on the CLEF-IP 2010/2011 collections to address some of the above challenges. The authors explored the scenario when only partial patent application is available for prior art search (e.g., *title, abstract, extended abstract*, or *description*). The authors tested different query expansion and reduction general methods such as [80] and a variant of the Maximal Marginal Relevance (MMR) [9]. They also tested patent-specific methods utilizing synonym sets [57], language models [23], and IPC-based lexicon [65]. After experimenting various sections as sources for the initial query terms as well as expansion/reduction sources, the results showed that the *description* section among other sections is the best to use as the initial query in the case of both query expansion and reduction. Query reduction was not beneficial for the long *description* queries as it already contains good coverage of relevant terms. However, query reduction on *description* queries was useful as it removed many of the noisy terms. Generally, query reduction outperformed query expansion on *description* and *extended abstract* queries which indicates that, with long queries, query reduction is effective for better retrieval performance. The results also showed that generic query expansion methods such as Rocchio works generally better for query expansion than patent-specific query expansion methods. Finally, the results showed that BM25 scoring works better than the TF-IDF scoring on the long *description* queries for both query reduction and expansion, while TF-IDF works better than BM25 on short and medium-length *title* or *abstract* queries. Through this comprehensive experimental study, the authors did not evaluate the impact of using multiple sections in combination as sources for query expansion or reduction. More importantly, the study does not provide any insights into the respective values of number of expansion terms or term removal threshold and whether these values are somewhat deterministic or vary widely calling for interactive setting.

To address the problem of poor PRF results in patent retrieval compared to traditional information retrieval, Bashir and Rauber [5] proposed a novel approach for PRF-based query expansion which builds a model that learns to identify better PRF results based on their similarity with the query patent over specific terms. These terms are learned by building a classification model that classifies whether a term would be useful for query expansion or not according to some proximity features between the original query terms and pseudo-relevant terms. The authors, through experiments on a subset of USPTO patents, showed the ability of this model to introduce more relevant query expansion terms and subsequently increasing the retrievability of individual patents. However, the authors did not evaluate this model on any of the available test collections. Moreover, extracting similarity features and computing similarities with PRF results during query execution are computationally expensive and time-consuming.

Along the same efforts, Mahdabi and Crestani [60] proposed a framework for identifying effective PRF documents at runtime and then performing query expansion using terms from these relevant documents. The authors first proposed patent-specific features and then used them to build a regression model which calculates a relevancy score of each PRF document. Though results on the CLEF-IP 2010 collection were encouraging, several challenges still exist, for example, the computational complexity of calculating the regression model features at runtime. And PRF parameters tuning (e.g., number of PRF documents to use).

Ganguly et al. [23] proposed a PRF approach which utilizes a language model for query reduction of long queries composed of full patent applications. The authors argued that naive application of PRF to expand query terms could add noisy terms causing query–topic drift. Moreover, naive removal of terms that has unit term frequency in the query could cause

removal of useful terms and thus hurt retrieval effectiveness. Instead, the authors proposed a PRF-based query reduction technique which generates language model similarity scores between query segments (sentences or n-grams) and top ranked results. Segments with top scores are kept and all others are removed. Results on the English subset of the CLEF-IP 2010 collection showed that the proposed approach outperforms the baselines. Parameter tuning is still the main downside of this technique. The performance of the proposed approach was unstable compared to the baselines with different parameter values, specifically the window size, the number of pseudo-relevant documents, and the fraction of terms to retain.

Golestan Far et al. [26] provided a study on hybrid QRE which aims to automatically approximate the optimal $\bar{Q}$ by careful selection/expansion of relevant query terms. To motivate the efficacy of QRE on retrieval performance, the authors first designed an experiment where relevance judgments of a query patent Q were assumed to be known in advance. After running Q, using PRF on top-k documents, only terms that are more frequent in retrieved relevant documents (those from relevance judgments) than irrelevant documents are kept and used as $\bar{Q}$. Then, querying using $\bar{Q}$ achieved a better performance than state of the art on the English subset of CLEF-IP 2010 collection. To approximate $\bar{Q}$ automatically, the authors proposed four different methods hoping to identify relevant vs. irrelevant terms in Q by: (1) removing terms with high document frequency in the top-100 retrieved documents, (2) removing infrequent terms in Q, (3) using frequent terms in relevant documents assuming the top-5 retrieved documents are relevant, and (4) performing query reduction on Q using IPC definitions as stop-words. All of the four methods failed to perform better than the keyword-based baseline. More interestingly, the authors demonstrated that baseline performance can be doubled if only one relevant document was manually provided by the user. This last observation motivates the need for interactive QRE as a simple and effective method for patent retrieval.

5.2.3 Semantic-based methods

As we mentioned before, in PR queries can vary from few terms (e.g., survey memo) to thousands of terms (e.g., full patent application). Straightforward keyword-based PR proved to be ineffective simply because of the vocabulary mismatch between query terms and relevant patents content. Magdy et al. [58] showed that, in the CLEF-IP 2009 collection, 12% of the relevant documents have no common words with the search topics. This motivates the need for novel approaches to bridge this vocabulary mismatch gap. Several semantic-based methods have been proposed in attempt to match queries with relevant documents based on their meanings rather than relying on keyword matches only. Table 5 shows some semantic-based methods along with their performance results on benchmark datasets.

Dictionary-based: semantic-based methods perform QRE by expanding the query to include other terms that have similar meanings to the original query terms. The first category of these methods are the dictionary-based techniques which use either generic [57], technical [48], or patent-specific dictionaries [62,88,89,91,92,97] for QRE. Generic dictionaries could be existing lexical databases such as WordNet [18], while patent-specific dictionaries are lexical databases generated from patent-related data such as examiner's query logs. In either case, similar or related terms to the original query terms are retrieved from such dictionaries and used for query expansion.

Magdy and Jones [57] explored the use of WordNet for query expansion in patent retrieval on the CLEF-IP 2010 collection. Overall, adding synonyms and hyponyms for nouns and verbs in the original query increased the MAP score slightly while decreased the PRES score significantly. Moreover, query execution time was increased considerably. The authors

Table 5 Semantic-based patent retrieval methods

Method	Description	Dataset	MAP	PRES
Magdy and Jones [57]	– Use Wordnet synonyms and hyponyms for query expansion	clef-ip 2010	0.136	0.484
	– Slow processing time		0.140⋆	0.486⋆
	– No improvement			
Tannebaum and Rauber [87] Tannebaum and Rauber [88] Tannebaum and Rauber [89] Tannebaum and Rauber [90] Tannebaum and Rauber [91] Tannebaum et al. [92]	– Mine query logs for synonyms, co-occurring, and proximity terms	clef-ip 2010	0.139	0.512
	– No improvement		0.139⋆	0.512⋆
	– Use upon request			
Magdy and Jones [57]	– Using synonyms learned from parallel translations (EN, GE, and FR)	clef-ip 2010	0.144	0.485
	– Improve MAP only		0.140⋆	0.486⋆
	– Use upon request			

⋆ Indicates baseline performance

considered this a "negative" result. As the use of WordNet was proven to be effective in other retrieval tasks [46,96], more experiments are needed to affirm the authors' conclusion, for example, investigating the impact of using synonyms only or hyponyms only and expanding terms belonging to specific sections or ambiguous terms only.

Recently, more research was focused on utilizing domain-specific and technical dictionaries rather than WordNet. Examiners' query logs have been an important resource for building such technical thesauri. Tannebaum and Rauber [87–91] and Tannebaum et al. [92] introduced an analysis of the USPTO examiners' search query logs. Their analysis, though on a subset of query logs, revealed interesting insights into patent examiners' search behavior which could be very useful for designing effective patent retrieval systems. For example, the authors noted about examiners' behavior while searching for prior art that: (1) the average query length is four terms, (2) search terms are mostly from the patent application under investigation, (3) expansion terms represent small percentage of query terms and mostly appear in the specific patent domain terminology, (4) the majority of query terms represents subject technical features that appears in the *claims* section, while very little percentage of them appears in the *description* section, (5) the majority of terms are nouns, followed by verbs, then adjectives, and (6) about half of the query operators used are "OR," followed by "AND," then proximity operators.

Tannebaum et al. built upon these insights and introduced methods to automatically identify synonyms/equivalents, co-occurring terms, and proximity relations for expanding query terms by mining examiners' search logs. As we can notice, learning expansion terms from query logs might be misleading because not all query sessions succeed to identify prior art. Additionally, deeper analysis of the query logs considering other metadata such as relevant hits count might be useful in this regard. On the other hand, it would be more useful if we can model the features of these terms, for example, based on their location, frequency, part of speech, etc. From effectiveness perspective, evaluating the generated lexical knowledge on the CLEF-IP 2010 collection did not record significant improvement [92]. Therefore, the authors recommended using it in an interactive mode rather than automatic mode to semi-automate query generation.

Corpus-based: the second category of semantic-based QRE is the corpus-based methods. In these methods, textual corpora are analyzed to extract semantically related concepts to query terms which can be used for query expansion. Al-Shboul and Myaeng [1] proposed a Wikipedia-based query expansion method which works by first creating a summary of each Wikipedia article containing the main category, all titles under the main category, and other categories with in/out links to the main category. At query time, query terms and phrases are matched with page summaries; then, phrases from matching pages are scored and selected for query expansion under the assumption that they are semantically related. Experiments on the subset of USPTO patents in the NTCIR-6 collection showed an increase in MAP over other query expansion techniques. However, the authors used IPC codes rather than citations as relevance judgments to topic queries which do not reflect the typical search practices, where it is needed to retrieve related patent documents not related to classification codes.

Another corpus-based method was proposed by Magdy and Jones [57], where synonym sets were automatically generated from the CLEF-IP patent corpus. The authors utilized parallel translations of patent sections in different languages to build a word-to-word translation model and infer synonymy relation when a word in one language is translated to multiple words in another language. These multiple words under some probabilistic threshold could be considered synonyms. Overall results using this method were better than PRF and Wordnet-based query expansion, but worse than the keyword-based baseline in Magdy and Jones [55]. The authors also showed that the performance of this method on some topics was better than the baseline which indicates its potential. The issue they raised is how to more effectively apply query expansion by selecting "good" terms [8], or predicting query expansion performance beforehand [13,54]. Such challenges can also be alleviated semi-automatically by developing intelligent and usable interactive query expansion frameworks which engage users in such decision. Finally, Krestel and Smyth [44] applied topic modeling of search hits in order to better rank retrieved patents. The results on a small collection of the USPTO patents showed improved MAP.

5.2.4 Metadata-based methods

Patents are not only textual documents, they contain lot of non-textual metadata and bibliographic information as well (e.g., citations, tables, formulas, drawings, classification, etc.). Combining metadata analysis with text-based PR has shown improvements in performance in the literature [16,19,48,49,61]. Metadata features are also language-independent making them advantageous when used for CLIR. Table 6 shows some metadata-based methods along with their performance results on benchmark datasets

Citation-based: The use of citation analysis for better patent retrieval is the most heavily reported technique of metadata-based methods. Naively incorporating citations from topic patent applications as prior art proved to be effective, eliminating the need for deeper citation analysis [59]. However, citation extraction from patent texts is challenging because there is no standard writing style for patent references. Lopez and Romary [49] developed a tool for citation mining which identifies, parses, normalizes, and consolidates patent citations. As citations might not be always available in all scenarios (e.g., related work search, technology survey), more mature techniques are needed. Fujii [19] proposed using PageRank [7] and document popularity as an additional scoring to re-rank query top results returned using *claims*-based queries. The results of applying popularity scoring on the English subset of NTCIR-6 improved MAP and recall over the raw text-based scoring. Incorporating PageRank, though intuitive, poses many challenges, especially because patent documents have references to the non-patent literature which would produce incomplete citation graph. Mahdabi and

Table 6 Metadata-based patent retrieval methods

Method	Description	Dataset	MAP	PRES
Fujii [19]	– Use PageRank on patents citation graph	ntcir-6	0.075	–
			0.081	
	– Use patent popularity among top results with weighted voting		0.071⋆	
Mahdabi and Crestani [61]	– Build query-specific citation graph from PRF results and their citations	clef-ip 2011	0.105	0.481
	– Weight nodes using PageRank		0.099⋆	0.450⋆
	– Estimate query LM from the graph nodes considering their PageRank scores			
Mahdabi and Crestani [63]	– Using time-aware random walk on weighted citation graph	clef-ip 2011	0.125	0.536
			0.058⋆	

⋆ Indicates baseline performance

Crestani [61] extended their query modeling technique in Mahdabi et al. [64] by incorporating term distributions of the PRF results as well as their citations in calculating the query language model. The authors first construct a query-specific citation graph using PRF results and their citations and assign a score for each of them using PageRank. Then, a query model is estimated from term distributions of the documents in the citation graph constrained by their respective PageRank. Finally, query expansion is performed using the estimated query model. Experiments on the CLEF-IP 2011 collection showed improved recall performance with no change in precision, which indicates the usefulness of using cited documents vocabulary for query expansion. Best improvements were achieved using the top 30 PRF documents, 2-levels citation graph, and 100 expansion terms. However, we can notice two main computational challenges using this technique in real-time setting: (1) computing the PageRank of the 2-level citations graph and (2) estimating the query model from top PRF documents as well as documents in the citation graph.

Classification-based: these methods utilize classification information of the topic patent and the retrieved documents to improve the performance of patent retrieval [11,24,32,33,42]. The naive use of IPC classification is to filter retrieved documents to keep only ones that share the same IPC classification code at some level (e.g., same subclass) with the topic patent [25,58]. The more sophisticated use of classification information was introduced by Verma and Varma [95] who proposed a new representation of patent documents based on IPC classifications. The method utilizes IPC codes assigned to the corpus patents as well as codes of their citing documents to form an IPC class vector. First, the vector is initialized from patent's IPC code, then codes of citing patents are propagated over multiple iterations. The most similar patents are retrieved using cosine similarity between IPC class vectors and re-ranked using text-based search utilizing the top 20 tf-idf topic patent terms. Experiments on the CLEF-IP 2011 collection showed improved recall but low MAP scores. The instability of the patent classification system poses a real challenge when it comes to

incorporating classification metadata into PR systems. Over time, new classes are added to the classification hierarchy and existing classes are expanded. In order to do reliable search based on classification codes, these changes must be accounted for periodically. Moreover, patents are assigned to multiple classification codes; however, almost all previous research considered only the primary class but not secondary classifications which might, if utilized, improve the retrieval performance.

Hybrid: these methods utilize various sources of metadata to improve PR performance. Mahdabi and Crestani [63] built upon previous work in Mahdabi et al. [64] and Mahdabi and Crestani [61] and proposed a query expansion method that utilizes time-aware random walk on a weighted patent citations network. Citation weights are derived from various metadata (e.g., classification codes, inventors, assignee, etc.). Citations with higher weights are considered more influential when performing query expansion. Experiments on the CLEF-IP 2010/2011 collections show improved recall and MAP. Mahdabi and Crestani [62] proposed building a query-specific lexicon from IPC definition pages and using it for query expansion. Unfortunately, the lexicon would be helpful only if the query represents a complete patent document with IPC codes assigned to it which is not always the case, especially at the early stages of the patent life cycle.

5.2.5 Interactive methods

Interactive patent retrieval is inevitable. As we can notice from the above review, effective fully automated retrieval of patent prior art is very challenging. Best methods perform around average in terms of PRES and much less in terms of MAP. Additionally, these methods require tuning a large number of parameters and thresholds whose optimal values differ according to the given query and the specific information need, for example, deciding which patent section to use, which PRF results, and which expansion terms and their respective weights. The answers of these questions are not deterministic and probably require multiple interaction cycles with the user in order to satisfy his/her information need.

Current interactive methods in patent retrieval are more focused on better organization, integration, and utilization of structured and textual patent data than on better retrieval performance. In other words, patent retrieval is addressed as a professional search problem rather than prior art search problem. Fafalios and Tzitzikas [17] presented a keyword-based interactive search framework to support patent search. The interaction elements are presented through post-analysis of search results in the form of facet-based features such as static metadata (e.g., IPC codes), textual clustering, named entity extraction, semantic enrichments, and others. The framework was applied on patent search [78] and evaluated using user study of twelve patent examiners [79]. Evaluation responses indicated overall acceptance of the framework in terms of usability, ease of use, efficiency, learnability. However, the authors did not report on the effectiveness or success of the system helping patent examiners to find prior art.

Shalaby and Zadrozny [83] proposed a visual interactive semantic framework for patent analysis which features semantic-based query expansion of search queries using Mined Semantic Analysis (MSA) [82]. In a nutshell, MSA builds an association knowledge graph using rule mining of concept rich textual corpora (e.g., Wikipedia). After mining the "See Also" link graph of Wikipedia, MSA could represent a topic query as a Bag of Concepts (BOC) derived from the association knowledge graph. This BOC could then be used to expand the original query terms. Figure 2 shows an example of the query expansion map of *Cognitive Analytics*. Another example is presented in Fig. 3 showing concept map of 10 patents of *Bank of America* using the *abstract* section. Users can interact with the concept

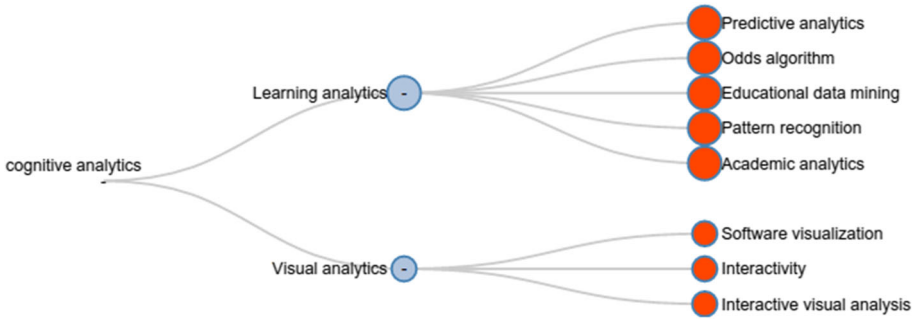

Fig. 2 Concept graph of cognitive analytics. Light blue nodes are explicit concepts, and red nodes are latent concepts (color figure online)

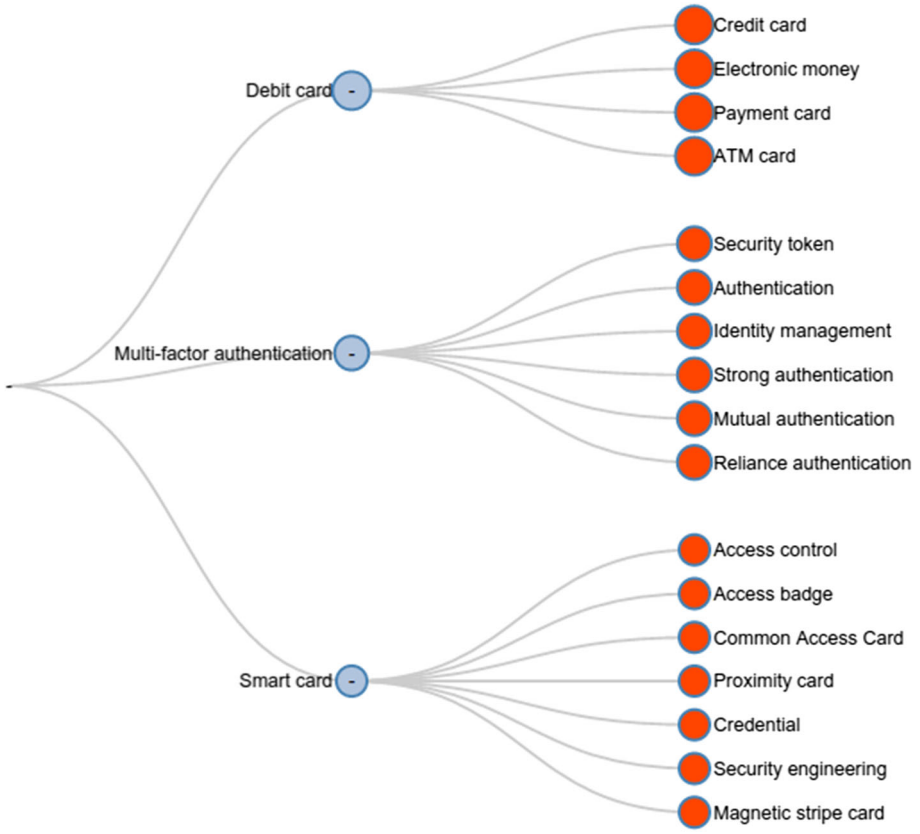

Fig. 3 Concept graph using Bank of America's 100 patent titles. Light blue nodes are explicit concepts, and red nodes are latent concepts (color figure online)

map by removing nodes and updating the search results. Shalaby et al. [84] demonstrated the applicability of their framework to support tasks such as prior art search, competitive intelligence, technology landscape analysis and exploration. However, they did not provide a controlled study evaluating the performance of their method on benchmark collections.

Developing interactive methods for patent retrieval is also motivated by recent analysis which showed significant performance improvement if only one relevant document was manually provided by the user [26]. Performance gains using technology-assisted review (TAR) [12,28] in domains such as electronic discovery motivates investigating the applicability of machine learning TAR protocols in patent retrieval.

Technology-assisted review, such as patent retrieval, is a total-recall task where it is required to find all relevant documents to the search request with reasonable effort (time and cost). It is thus a human-in-the-loop process where a human expert manually annotates a subset of the documents as relevant or irrelevant. The underlying algorithm subsequently builds a ranking model by training on such annotations and uses this model to promote more relevant results and demote irrelevant ones as more documents are searched and annotated. This process stops when enough results are obtained. Typically, these algorithms utilize techniques such as continuous active learning combined with Boolean search in order to develop and adapt the ranking model [28].

Several questions still need to be addressed when it comes to investigating technology-assisted review protocols applicability to patent retrieval, as these protocols were only evaluated in ad hoc search scenarios. The complexity of patents terminology and availability of multiple sources of metadata would, likely, demonstrate many opportunities for adaptation and modifications to the current technology-assisted review protocols.

6 Related topics

Despite intense interest within the research community in patent retrieval, the patent industry has many other challenges and open problems which are of high interest and value to various stakeholders, such as economists, R&D managers, and legal professionals, to name a few. In this section, we try to lightly touch on these tasks and highlight some challenges and possible future directions.

6.1 Patent quality assessment

Assessing the technical quality and importance of inventions is very important to patent owners because it allows them to:

- Better utilize their IP management costs by automated recommendation of patent maintenance decisions.
- Better determine the novelty and originality of their patents.
- Maximize licensing revenues by automatic estimation of the patent value.

Because there is no ground truth for quality measurements, performance evaluation of quality assessment techniques is usually based on indicators such as correlation with patent forward citations, maintenance status history, court rulings (if any), and/or patent reexamination history (if any). Some early work scored patents using their metadata such as citations count, maintenance history, global prosecution efforts [45], and even manually by patent attorneys. Automated patent quality assessment has gained more traction in recent years though.

Citation analysis has been and still a main technique for patent valuation [14,29,31,93,100]. Wang et al. [100] proposed a probabilistic mixture approach to predict whether a topic patent will be renewed at different renewal periods. The method first divides the citations into two groups, technological and legal. From each group, different features reflecting the technological richness, technological influence, legal patent scope, and legal blocking power of each patent are combined. The authors subsequently built a binary classifier using these probabilistic features. Evaluation is performed by comparing the model's predictions against the renewal decisions of a collection of patents. While proved effective, estimating patent value as a binary outcome might not be practical, especially if a patent owner needs to prioritize his maintenance decisions of multiple patents.

Quality assessment based on the lexical features of the patent text was also explored in the literature [35,40,47]. Liu et al. [47] proposed a graphical model to estimate patent quality as a latent variable. The model utilized lexical features extracted from the patent text such as *claims* n-grams age and popularity, lexical alignment between the *claims* and the *description*, number of dependent and independent claims, number of reported classes when filing the patent, and other features. The authors also incorporated measurements such as forward citations count, court decisions, and reexamination records. It is clear that court decisions are only available for small number of patents which might not allow building a robust model.

Jin et al. [40] modeled the patent maintenance decision as recommendation problem where patents were represented as multi-modal heterogeneous information network. The model utilized several metadata features, lexical features such as unique words and lengths of different sections, as well as inventor and assignee profile features. Experimental results showed high prediction accuracy on a large number of USPTO patents.

Hu et al. [37] proposed a time-based topic model which ranks patents novelty and influence based on whether the dominant topics in patent's prior art (for novelty) or forward art (for influence) are still active topics. The authors also proposed using time decay function to address the problem of old patents having less prior art and more forward art than newer patents and vice versa. Results showed high correlation between assigned ranks and forward citations count.

Hido et al. [35] proposed a scoring model which assigned a patentability score to each patent and thus can be utilized to determine whether it will be granted. First, the authors extracted textual features such as word frequency, word age, and syntactic complexity (e.g., number of sentences). Then, they trained a classifier using previous patent office decisions as ground truth. Though results showed the model effectiveness, the utilized syntactic complexity features are all extracted from the topic patent and thus could be good predictors for the writing quality not patentability potential.

The correlation between patent *claims* novelty and patent value using lexical analysis of patent text has been analyzed in previous studies [10,34]. Hasan et al. [34] proposed an IR-based ranking tool which analyzes patent *claims* for originality. The technique first extracts key terms and phrases from the *claims* text using syntactic patterns and then looks for usage patterns backward to determine their novelty, and forward to determine their influence. The method considers usage patterns only through user-defined time window. It is also keyword-based and hence will fail to capture key phrases that are semantically similar and subsequently might give inaccurate scores.

Along the efforts of using patent legal data for quality assessment, Mann and Underweiser [66] utilized prosecution histories, court decisions, and patent textual features to analyze patent quality. The analysis suggested that patent examination records would be very helpful in better discriminating high-quality patents from low-quality patents and possibly improve the examination process as a whole.

6.2 Patent litigation

Litigation in general, and patent litigation specifically, has been and still a topic of interest to legal professionals. With the increased amounts of digitized data available and the need for technology support in analyzing and mining these huge datasets, litigation became of more interest to computational science researchers. Patent litigation can take many forms, the most common is patent infringement litigation where a patent owner (plaintiff) accuses another party (defendant) of using his/her invention without license or permission. Because litigation is very expensive, the most common defensive action for the defendant is to establish invalidity of the plaintiff invention by issuing a post-grant proceeding such as post-grant review or inter-parts review. Now the problem becomes a patent retrieval task, i.e., invalidity search, where one of the aforementioned methods can be utilized with wider scope to cover not only the patent literature but also other published materials.

The task of automatically establishing patent infringement is not addressed in the literature. Such task requires extensive human expertise and reasoning to build correspondences between product features and patent claims. On the other hand, statistical and visual analytics of previous court decisions have shown some degree of success in helping lawyers to better understand possible outcomes and better plan on defense strategies [2,30,69].

For example, Allison et al. [3] provided a statistical study on patent cases filed from 2008 to 2009 and decisions made between (2009 and 2013). The study showed that there is a strong correlation between court decision, and patent-specific, litigation-specific, and industry-specific variables such as industry and technology type, inventors foreign status, number of claims, number of forward and backward citations, and number of defendants sued.

Rajshekhar et al. [75] studied the potential of concept-based semantic search in patent litigation. The authors designed an experiment in order to retrieve invalidating patents to a given litigated patent using a subset of PTAB's final decisions as ground truth and a search corpus of $\sim$7 m USPTO patents. The authors, based on the experimental results and through interviews with patent practitioners, concluded that a one-size-fits-all semantic search approach is incapable of capturing the highly nuanced relevance judgments made in the domain of patent litigation. Rather, the search workflow should be modeled as a multistage information seeking process, where users are presented with interactive elements to control the search space, and their feedback is incorporated iteratively in the relevance ranking of retrieved results for enhanced performance.

Finally, There is much to be done in building predictive models for patent litigation given the availability of prior case datasets that were not available few years ago (e.g., prosecution histories, court decisions, and PTAB decisions).

6.3 Technology licensing

Patents represent one of the most valuable assets in today's enterprises which, if leveraged effectively, guarantee not only competitive superiority, but also huge licensing revenues [10]. The technology licensing task is three-sided. First, patent owners would be interested in finding potential licensees with reasonable effort. Second, licensees would like to find relevant inventions to their businesses. Third, owners and businesses would be interested in gauging the strategic and protection values of a patent in order to support their pricing and offering decisions.

While there is not much research focusing on automatically recommending potential licensees, the task of recommending patents to be licensed was relatively more considered. Chen et al. [10] proposed a platform called SIMPLE which is used at IBM to identify target patents for licensing. Given a set of topic patents, SIMPLE uses nearest neighbor similarity to find other patents that are most similar to the given topic set. Then, all the patents are grouped and proposed as one licensing package to interested party. The platform was extended in Spangler et al. [85] to allow retrieving target patents using free-text search. We can notice that current trends for identifying potential patents for licensing model the problem as a PR task. More elaboration on the SIMPLE platform was introduced by Spangler et al. [86] using interactive visualization. First, portfolios of two companies are contrasted to find content overlap between both of them using proximal search. Then, the closest patents to the overlap area are recommended as candidates for licensing.

7 Concluding remarks

In this paper, we presented a comprehensive review of patent retrieval methods and approaches. It is clear that the well-performing information retrieval techniques in areas such as Web search cannot be utilized directly in PR without deliberate domain adaptation and customization. Furthermore, state-of-the-art performance in automatic patent retrieval is still low (< 0.2 MAP). Several proposed techniques for query expansion, query reduction and pseudo-relevance feedback require tuning of various parameters. Professional search practices suggest that effective prior art search requires multiple iterations of searching, reviewing, and refining. On the other hand, examiners' query formulation practices (few keywords and Boolean search) are different from those of automatic methods (many keywords and free-text search). These observations motivate the need for interactive search tools which provide cognitive assistance to search professionals with minimal effort. These tools must also be developed in hand with patent professionals considering their practices and expectations.

Unexplored patent-related data sources might be an opportunity for breakthrough improvements over the current modest state of the art in patent retrieval, for example, utilizing reexamination records, PTAB decisions, differences between the patent application and the granted version, examiner/applicant correspondences, and prosecution histories. All these resources are not yet fully explored in the literature of patent retrieval.

Related tasks such as patent quality assessment, litigation, and licensing are of less focus among computational scientists. However, they provide wide opportunities for future exploration from computational and modeling perspectives. These tasks require interdisciplinary and cooperative efforts from both legal professionals and the computer science research community.

Acknowledgements This work was supported by the National Science Foundation (Grant No. 1624035). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Al-Shboul B, Myaeng SH (2014) Wikipedia-based query phrase expansion in patent class search. Inf Retr 17(5–6):430–451
2. Allison JR, Lemley MA, Schwartz DL (2013) Understanding the realities of modern patent litigation. Tex L Rev 92:1769

3. Allison JR, Lemley MA, Schwartz DL (2015) Our divided patent system. Univ Chic Law Rev 82(3):1073–1154
4. Baeza-Yates R, Ribeiro-Neto B et al (1999) Modern information retrieval, vol 463. ACM Press, New York
5. Bashir S, Rauber A (2010) Improving retrievability of patents in prior-art search. In: Advances in information retrieval, Springer, pp 457–470
6. Bouadjenek MR, Sanner S, Ferraro G (2015) A study of query reformulation for patent prior art search with partial patent applications. In: Proceedings of the 15th international conference on artificial intelligence and law, ACM, pp 23–32
7. Brin S, Page L (2012) Reprint of: the anatomy of a large-scale hypertextual web search engine. Comput Netw 56(18):3825–3833
8. Cao G, Nie JY, Gao J, Robertson S (2008) Selecting good expansion terms for pseudo-relevance feedback. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 243–250
9. Carbonell J, Goldstein J (1998) The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 335–336
10. Chen Y, Spangler S, Kreulen J, Boyer S, Griffin TD, Alba A, Behal A, He B, Kato L, Lelescu A, et al (2009) Simple: a strategic information mining platform for licensing and execution. In: IEEE international conference on data mining workshops, 2009. ICDMW'09, IEEE, pp 270–275
11. Chen YL, Chiu YT (2011) An IPC-based vector space model for patent retrieval. Inf Process Manag 47(3):309–322
12. Cormack GV, Grossman MR (2014) Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In: Proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval, ACM, pp 153–162
13. Cronen-Townsend S, Zhou Y, Croft WB (2002) Predicting query performance. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 299–306
14. Czarnitzki D, Hussinger K, Leten B (2011) The market value of blocking patent citations. ZEW - Zentrum für Europäische Wirtschaftsforschung/Center for European Economic Research
15. D'hondt E, Verberne S (2010) Clef-ip 2010: Prior art retrieval using the different sections in patent documents. In: CLEF (Notebook Papers/LABs/Workshops)
16. Eisinger D, Tsatsaronis G, Bundschus M, Wieneke U, Schroeder M (2013) Automated patent categorization and guided patent search using IPC as inspired by mesh and pubmed. J Biomed Semant 4(1):1
17. Fafalios P, Tzitzikas Y (2014) Exploratory professional search through semantic post-analysis of search results. In: Professional search in the modern world, Springer, pp 166–192
18. Fellbaum C (1998) WordNet: an electronic lexical database. Bradford Books, Cambridge
19. Fujii A (2007) Enhancing patent retrieval by citation analysis. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 793–794
20. Fujii A, Iwayama M, Kando N (2004) Overview of patent retrieval task at ntcir-4. In: NTCIR
21. Fujii A, Iwayama M, Kando N (2005) Overview of patent retrieval task at ntcir-5. In: In Proceedings of the fifth NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access, pp 269–277
22. Fujii A, Iwayama M, Kando N (2007) Overview of the patent retrieval task at the ntcir-6 workshop. In: NTCIR
23. Ganguly D, Leveling J, Magdy W, Jones GJ (2011) Patent query reduction using pseudo relevance feedback. In: Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, pp 1953–1956
24. Giachanou A, Salampasis M, Paltoglou G (2015) Multilayer source selection as a tool for supporting patent search and classification. Inf Retr J 18(6):559–585
25. Gobeill J, Pasche E, Teodoro D, Ruch P (2009) Simple pre and post processing strategies for patent searching in CLEF intellectual property track 2009. In: Multilingual information access evaluation I: text retrieval experiments, Springer, pp 444–451
26. Golestan Far M, Sanne S, Bouadjenek MR, Ferraro G, Hawking D (2015) On term selection techniques for patent prior art search. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, ACM, pp 803–806
27. Graf E, Azzopardi L (2008) A methodology for building a patent test collection for prior art search. In: Proceedings of the 2nd international workshop on evaluating information access, EVIA

28. Grossman MR, Cormack GV (2011) Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. Rich JL & Tech 17:11–16
29. Hall BH, Jaffe A, Trajtenberg M (2005) Market value and patent citations. RAND J Econ 36(1):16–38
30. Harbert T (2013) The law machine. Spectrum 50(11):31–54
31. Harhoff D, Narin F, Scherer FM, Vopel K (1999) Citation frequency and the value of patented inventions. Rev Econ Stat 81(3):511–515
32. Harris CG, Foster S, Arens R, Srinivasan P (2009) On the role of classification in patent invalidity searches. In: Proceedings of the 2nd international workshop on patent information retrieval, ACM, pp 29–32
33. Harris CG, Arens R, Srinivasan P (2010) Comparison of ipc and uspc classification systems in patent prior art searches. In: Proceedings of the 3rd international workshop on patent information retrieval, ACM, pp 27–32
34. Hasan MA, Spangler WS, Griffin T, Alba A (2009) COA: finding novel patents through text analysis. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 1175–1184
35. Hido S, Suzuki S, Nishiyama R, Imamichi T, Takahashi R, Nasukawa T, Idé T, Kanehira Y, Yohda R, Ueno T et al (2012) Modeling patent quality: a system for large-scale patentability analysis using text mining. Inf Med Technol 7(3):1180–1191
36. Hiemstra D, Robertson S, Zaragoza H (2004) Parsimonious language models for information retrieval. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, ACM, pp 178–185
37. Hu P, Huang M, Xu P, Li W, Usadi AK, Zhu X (2012) Finding nuggets in IP portfolios: core patent mining through textual temporal analysis. In: Proceedings of the 21st ACM international conference on Information and knowledge management, ACM, pp 1819–1823
38. Iwayama M, Fujii A, Kando N, Takano A (2003) Overview of patent retrieval task at NTCIR-3. In: Proceedings of the ACL-2003 workshop on Patent corpus processing, vol 20, association for computational linguistics, pp 24–32
39. Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst (TOIS) 20(4):422–446
40. Jin X, Spangler S, Chen Y, Cai K, Ma R, Zhang L, Wu X, Han J (2011) Patent maintenance recommendation with patent information network model. In: 2011 IEEE 11th international conference on data mining (ICDM), IEEE, pp 280–289
41. Jürgens JJ, Hansen P, Womser-Hacker C (2012) Going beyond CLEF-IP: the reality for patent searchers? In: Information access evaluation. Multilinguality, multimodality, and visual analytics, Springer, pp 30–35
42. Kim J, Kang IS, Lee JH (2006) Cluster-based patent retrieval using international patent classification system. In: Computer processing of oriental languages. Beyond the orient, the research challenges ahead, Springer, pp 205–212
43. Konishi K (2005) Query terms extraction from patent document for invalidity search. In: NTCIR
44. Krestel R, Smyth P (2013) Recommending patents based on latent topics. In: Proceedings of the 7th ACM conference on Recommender systems, ACM, pp 395–398
45. Lanjouw JO, Pakes A, Putnam J (1998) How to count patents and value intellectual property: the uses of patent renewal and application data. J Ind Econ 46(4):405–432
46. Liu S, Liu F, Yu C, Meng W (2004) An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 266–272
47. Liu Y, Hseuh Py, Lawrence R, Meliksetian S, Perlich C, Veen A (2011) Latent graphical models for quantifying and predicting patent quality. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 1145–1153
48. Lopez P, Romary L (2009) Multiple retrieval models and regression models for prior art search. In: CLEF 2009 workshop
49. Lopez P, Romary L (2010) Experiments with citation mining and key-term extraction for prior art search. In: CLEF 2010-conference on multilingual and multimodal information access evaluation
50. Lupu M, Huang J, Zhu J, Tait J (2009) Trec-chem: large scale chemical information retrieval evaluation at trec. In: ACM SIGIR forum, ACM, vol 43, pp 63–70
51. Lupu M, Tait J, Huang J, Zhu J (2010) Trec-chem 2010: notebook report. Proc TREC 2010:2
52. Lupu M, Mayer K, Tait J, Trippe AJ (2011a) Current challenges in patent information retrieval, vol 29. Springer, Berlin
53. Lupu M, Zhao J, Huang J, Gurulingappa H, Fluck J, Zimmermann M, Filippov IV, Tait J (2011b) Overview of the TREC 2011 chemical IR track. In: TREC

54. Lv Y, Zhai C (2009) Adaptive relevance feedback in information retrieval. In: Proceedings of the 18th ACM conference on Information and knowledge management, ACM, pp 255–264
55. Magdy W, Jones GJF (2010) Applying the KISS principle for the CLEF-IP 2010 prior art candidate patent search task. Dublin City University, CLEF labs
56. Magdy W, Jones GJ (2010b) 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, ACM, pp 611–618
57. Magdy W, Jones GJ (2011) A study on query expansion methods for patent retrieval. In: Proceedings of the 4th workshop on Patent information retrieval, ACM, pp 19–24
58. Magdy W, Leveling J, Jones GJ (2009) Exploring structured documents and query formulation techniques for patent retrieval. In: Multilingual information access evaluation I: text retrieval experiments, Springer, pp 410–417
59. Magdy W, Lopez P, Jones GJ (2011) Simple vs. sophisticated approaches for patent prior-art search. In: Advances in information retrieval, Springer, pp 725–728
60. Mahdabi P, Crestani F (2012) Learning-based pseudo-relevance feedback for patent retrieval. In: Multidisciplinary information retrieval, Springer, pp 1–11
61. Mahdabi P, Crestani F (2014a) The effect of citation analysis on query expansion for patent retrieval. Inf Retr 17(5–6):412–429
62. Mahdabi P, Crestani F (2014b) Patent query formulation by synthesizing multiple sources of relevance evidence. ACM Trans Inf Syst (TOIS) 32(4):16
63. Mahdabi P, Crestani F (2014c) Query-driven mining of citation networks for patent citation retrieval and recommendation. In: Proceedings of the 23rd ACM International conference on information and knowledge management, ACM, pp 1659–1668
64. Mahdabi P, Keikha M, Gerani S, Landoni M, Crestani F (2011) Building queries for prior-art search. Springer, Berlin
65. Mahdabi P, Gerani S, Huang JX, Crestani F (2013) Leveraging conceptual lexicon: query disambiguation using proximity information for patent retrieval. In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval, ACM, pp 113–122
66. Mann RJ, Underweiser M (2012) A new look at patent quality: relating patent prosecution to validity. J Empir Leg Stud 9(1):1–32
67. Meij E, Weerkamp W, de Rijke M (2009) A query model based on normalized log-likelihood. In: Proceedings of the 18th ACM conference on Information and knowledge management, ACM, pp 1903–1906
68. NTCIR (2015) NTCIR test collections. http://research.nii.ac.jp/ntcir/permission/data-en.htm, http://research.nii.ac.jp/ntcir/permission/data-en.htm. Accessed 30 Apr 2016
69. Osbeck MK (2015) Using data analytics tools to supplement traditional research and analysis in forecasting case outcomes. U of Michigan Public Law Research Paper Series (446)
70. Osborn M, Strzalkowski T, Marinescu M (1997) Evaluating document retrieval in patent database: a preliminary report. In: Proceedings of the sixth international conference on Information and knowledge management, ACM, pp 216–221
71. Piroi F, Lupu M, Hanbury A, Sexton AP, Magdy W, Filippov IV (2010) CLEF-IP 2010: retrieval experiments in the intellectual property domain. In: CLEF (notebook papers/labs/workshops)
72. Piroi F, Lupu M, Hanbury A, Zenz V (2011) CLEF-IP 2011: retrieval in the intellectual property domain. In: CLEF (notebook papers/labs/workshop), Citeseer
73. Piroi F, Lupu M, Hanbury A, Magdy W, Sexton A, Filippov I (2012) CLEF-IP 2012: retrieval experiments in the intellectual property domain, vol 1178, CEUR-WS
74. Piroi F, Lupu M, Hanbury A (2013) Information access evaluation. In: Proceedings of CLEF 2013 4th international conference of the CLEF initiative multilinguality, multimodality, and visualization, Valencia, Spain, September 23–26, 2013, Springer, Berlin, chap Overview of CLEF-IP 2013 Lab, pp 232–249. https://doi.org/10.1007/978-3-642-40802-1_25
75. Rajshekhar K, Shalaby W, Zadrozny W (2016) Analytics in post-grant patent review: possibilities and challenges (preliminary report). In: Proceedings of the American society for engineering management 2016 international annual conference
76. Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M et al (1995) Okapi at trec-3. NIST Spec Publ SP 109:109
77. Roda G, Tait J, Piroi F, Zenz V (2010) Multilingual information access evaluation I. Text retrieval experiments: 10th workshop of the cross-language evaluation forum, CLEF 2009, Corfu, Greece, September 30–October 2, 2009, Revised selected papers, Springer, Berlin, chap CLEF-IP 2009: retrieval experiments in the intellectual property domain, pp 385–409. https://doi.org/10.1007/978-3-642-15754-7_47

78. Salampasis M, Hanbury A (2014) Perfedpat: an integrated federated system for patent search. World Pat Inf 38:4–11
79. Salampasis M, Giachanou A, Hanbury A (2014) An evaluation of an interactive federated patent search system. In: Multidisciplinary information retrieval, Springer, pp 120–131
80. Salton G (1971) The SMART retrieval system-experiments in automatic document processing. Prentice-Hall Inc, Upper Saddle River
81. Schwartz DL, Sichelman TM (2015) Data sources on patents, copyrights, trademarks, and other intellectual property. Copyrights, Trademarks, and Other Intellectual Property (August 17, 2015) 2
82. Shalaby W, Zadrozny W (2015) Measuring semantic relatedness using mined semantic analysis. arXiv preprint arXiv:1512.03465
83. Shalaby W, Zadrozny W (2016) Innovation analytics using mined semantic analysis. In: Proceedings of the 29th international FLAIRS conference
84. Shalaby W, Rajshekhar K, Zadrozny W (2016) A visual semantic framework for innovation analytics. In: Proceedings of the thirtieth AAAI conference on artificial intelligence (AAAI-16). http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12303/12306
85. Spangler S, Chen Y, Kreulen J, Boyer S, Griffin T, Alba A, Kato L, Lelescu A, Yan S (2010) Simple: interactive analytics on patent data. In: 2010 IEEE international conference on data mining workshops (ICDMW), IEEE, pp 426–433
86. Spangler S, Ying C, Kreulen J, Boyer S, Griffin T, Alba A, Kato L, Lelescu A, Yan S (2011) Exploratory analytics on patent data sets using the simple platform. World Pat Inf 33(4):328–339
87. Tannebaum W, Rauber A (2012a) Acquiring lexical knowledge from query logs for query expansion in patent searching. In: 2012 IEEE sixth international conference on semantic computing (ICSC), IEEE, pp 336–338
88. Tannebaum W, Rauber A (2012b) Analyzing query logs of USPTO examiners to identify useful query terms in patent documents for query expansion in patent searching: a preliminary study. In: Multidisciplinary information retrieval, Springer, pp 127–136
89. Tannebaum W, Rauber A (2013) Mining query logs of uspto patent examiners. In: Information access evaluation. Multilinguality, multimodality, and visualization, Springer, pp 136–142
90. Tannebaum W, Rauber A (2014) Using query logs of uspto patent examiners for automatic query expansion in patent searching. Inf Retr 17(5–6):452–470
91. Tannebaum W, Rauber A (2015) Patnet: a lexical database for the patent domain. In: Advances in information retrieval, Springer, pp 550–555
92. Tannebaum W, Mahdabi P, Rauber A (2015) Effect of log-based query term expansion on retrieval effectiveness in patent searching. In: Experimental IR meets multilinguality, multimodality, and interaction, Springer, pp 300–305
93. Trajtenberg M (1990) A penny for your quotes: patent citations and the value of innovations. Rand J Econ 21(1):172–187
94. Verberne S, D'hondt E (2009) Prior art retrieval using the claims section as a bag of words. In: Multilingual information access evaluation I: text retrieval experiments. Springer, pp 497–501
95. Verma M, Varma V (2011) Patent search using IPC classification vectors. In: Proceedings of the 4th workshop on patent information retrieval, ACM, pp 9–12
96. Voorhees EM (1998) Using wordnet for text retrieval. Fellbaum (Fellbaum, 1998) pp 285–303
97. Wajda J, Zadrozny W (2016) Challenging problems and solutions in intelligent systems. In: Chap prior-art relevance ranking based on the examiner's query log content, Springer International Publishing, Cham, pp 323–333. https://doi.org/10.1007/978-3-319-30165-5_15
98. Wanagiri MZ, Adriani M (2010) Prior art retrieval using various patent document fields contents. In: CLEF (Notebook Papers/LABs/Workshops)
99. Wang F, Lin L (2015) Query construction based on concept importance for effective patent retrieval. In: 2015 12th international conference on fuzzy systems and knowledge discovery (FSKD), IEEE, pp 1455–1459
100. Wang S, Lei Z, Lee WC (2014) Exploring legal patent citations for patent valuation. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management, ACM, pp 1379–1388
101. Xue X, Croft WB (2009) Transforming patents into prior-art queries. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, ACM, pp 808–809

Walid Shalaby Walid Shalaby is currently a Research Scientist at the Industrial AI Lab, Hitachi America R&D. He received his PhD in Computer Science from the University of North Carolina at Charlotte in 2018. Dr. Walid's research interests include data & text mining, knowledge extraction, information retrieval, and machine learning. His PhD research involved developing methods for concept-based semantic representations and understanding of natural languages addressing tasks such as semantic search, entity-type recognition, and neural-based representations of concepts and entities. His research also involved enhancing the performance of deep learning architectures through unsupervised pre-training and entity awareness. Dr. Walid developed mined semantic analysis (MSA), a patented technique for concept-based text representation. Dr. Walid has several years of research and industry experience; through three internships at CareerBuilder and Samsung Research America, he worked on designing algorithms, developing prototypes, and modeling of data-driven solutions for systems such as personal digital assistants, job recommendation engines, and large-scale job search engines. Dr. Walid received his Bachelor's and Master's degrees in Computer Science from the Faculty of Computers and Information, Cairo University.

Wlodek Zadrozny Wlodek Zadrozny joined the faculty of the University of North Carolina in Charlotte in 2013, after a 27-year career at IBM T.J. Watson Research Center. Dr. Zadrozny's research focuses on natural language understanding and its applications. From 2008 to 2013, he was part of the IBM Watson project—the Jeopardy! playing machine. As a scientist at IBM Research, he led and contributed to a wide range of projects. They included semantic search applications, natural language dialogue systems, and a value net analysis of intangible assets. Dr. Zadrozny published over seventy refereed papers on various aspects of text processing; he is an author of over fifty patents granted and several patents pending. Wlodek Zadrozny received a PhD in Mathematics (with distinction) from Polish Academy of Science in 1980.