

A methodology for building a patent test collection for prior art search

Erik Graf
University of Glasgow
graf@dcs.gla.ac.uk

Leif Azzopardi
University of Glasgow
leif@dcs.gla.ac.uk

Abstract

This paper proposes a methodology for the construction of a patent test collection for the task of prior art search. Key to the justification of the methodology is an analysis of the nature and structure of patent documents and the patenting process. These factors enable a corpus of patent documents to be reverse engineered in order to arrive at high quality, realistic, relevance assessments. The paper first outlines the case for such a prior art search test collection along with the characteristics of patent documents, before describing the proposed method. Further research and development will be directed towards the application of this methodology to create a suite of prior art search topics for the evaluation of patent retrieval systems. We also include a preliminary analysis of its application on European patents.

Keywords: *Prior Art Search, Patent Retrieval, Test Collection, Evaluation.*

1 Introduction

Test collections play a vital role in the evaluation of retrieval systems [37, 38, 46]. Existing collections have enabled IR research to be conducted on the retrieval of news stories, web pages and government documents. Other areas actively being explored are blogs and enterprise documents. One area that provides a distinctly different set of research problems are patents. Patent documents can be of great value; and can have a major economic impact [21, 36]. For example, the European Patent Office ‘estimates that European industry is losing US\$20 billion every year due to lack of patent information, which results in duplication of effort such as re-inventing existing inventions, resolving problems that have already been solved, and redeveloping products that already are on the market’ [22]. Moreover patents are an invaluable source of scientific and technological information and are common subjects of study in scientific areas such as economics [7, 15], scientometrics [40, 48], and law [33]. Thus, it is important to conduct research into patent retrieval. While the need to conduct patent retrieval

research has been long recognized [31, 42], there has been a lack of test collections available.

One of the reasons for the lack of patent retrieval test collections stems from the complexity introduced by the dual nature of patents. Patents are devised as means of intellectual property protection, and exhibit both informative and judicial characteristics. Determining the relevance of a patent document is therefore a task requiring legal as well as subject expertise. This renders the task of test collection creation significantly more difficult than in other domains, such as web pages or news stories.

In this paper, we propose a method which can be applied in order to create topics for prior art search, the task of identifying all information that might be relevant to a patent’s claims of novelty. Our method exploits the process in which a patent document is created, in order to infer relevance assessments for prior art search topics. The main advantage of the methodology lies in its ease of application to different patent collections in order to create numerous topics with high quality, realistic assessments, without any recourse to any further specific subject or legal expertise.

The remainder of this paper is structured in the following way: Section 2 will explore related work. In Section 3 we outline the concept of test collections, the role they are playing in IR evaluation, and illustrate and discuss the central elements of such a collection. As the patent domain significantly differs in many aspects Section 4 will provide a brief introduction to the patent system and discuss characteristics relevant to the creation of a test collection. In Section 5 we introduce our proposed methodology for creating a prior art test collection. Section 6 explores the applicability of the methodology on European Patent Office (EPO) patents. In the final section we will discuss our findings and provide an outlook to further research.

2 Background

Since their introduction, test collections in Information Retrieval have played a pivotal role in the evaluation of retrieval models. One of the first test collections was defined as part of the Cranfield Experiments

[12] and provided the blue print for subsequent IR test collections. The ‘significant achievement of Cranfield 2 was to define a notion of the methodology of IR experimentation’ [37]. A design goal for the Cranfield 2 experiment was to create a laboratory type situation by reducing the number of operational variables during experiments.

Although not unchallenged [29], this approach has found widespread adoption in IR, and can nowadays be seen as the standard system evaluation method. Moreover a number of evaluation forums have been created, that adopted and extended this prior work for their deployed test collections. Most notably since its inception in 1992, the Text REtrieval Conference (TREC) [3] held by the U.S. National Institute of Standards and Technology (NIST), has provided an invaluable test bed for IR evaluation. TREC has produced many test collections covering a variety of tasks in a number of domains, such as the World Wide Web (WWW), legal, government, blogs, and enterprise. Predominately, TREC has provided collections based on English documents. Other forums include the Cross Language Evaluation Forum (CLEF) [1] which provides a comparable set of test collections in a number of European languages, and NII Test Collections for IR Systems project (NTCIR) [2] with a focus on East Asian languages. It is only the latter of these forums that has seriously considered patent documents. Subsequently we will provide an overview of the patent related collections and tasks deployed at it.

First introduced in the third NTCIR workshop, the patent task has led to the release of several patent test collections. Details of these collections are provided in Table 1. These test collections, primarily targeted at Japanese patent documents, have been associated with a variety of different user tasks. The listing below provides an overview of these tasks.

1. Cross language, cross genre retrieval (NTCIR 3 [27]): Given Japanese, English, and Chinese newspaper articles associated with a technology or commercial products, the task consisted of retrieving Japanese patents relevant to the article. Assessments for this task were conducted manually.
2. Associative retrieval (NTCIR 3) [27]: The task consists of retrieving patents for a given search topic (i.e. either a newspaper article or patent). Participants were asked to submit a list of retrieved patents and passages associated with the topic.
3. Invalidity search (NTCIR 4 [18],5 [19],6 [20]): Participants were asked to search a target patent collection for patents that can invalidate the demand in a given claim. In practice, for each search topic (i.e. a claim), each group submits

a list of retrieved patents and passages associated with the topic. The task was aiming at identifying patents that can invalidate a topic claim by themselves (1) or in combination with other patents (2).

4. Patent classification (NTCIR 5 [24],6 [25]): The purpose of this task lies in categorizing target patent applications based on the F-term classification system. A submission consisted of a ranked list of F-term classification codes for each target patent application.

The described tasks and collections provide a significant step towards patent retrieval specifically in Japanese. Our work differs from this work in its focus on the prior art task, and our aim to develop an underlying methodology that can be applied to a variety of patent sources in different languages. However, valuable lessons learnt at NTCIR can be taken on board to develop our methodology. The idea of using inferred relevance assessments, as done at NTCIR 5 and 6, will also form the method of creation of Prior Art search topics employed in our methodology.

3 Test Collections

To allow for the measurement of information retrieval effectiveness in a standardized way a test collection consists of three elements: A document collection (corpus), a task represented by a suite of specified information needs (topics), and a set of relevance judgments associated with the topics. The remainder of this section outlines necessary considerations for each of these elements on designing a test collection. Generally in order to avoid topic specific bias, the document collection and suite of information needs have to be of a reasonable size to enable averaging performance over topics and documents [47].

1. **Corpus:** The set of documents comprising the collection, generally should be a ‘sample of the kinds of texts that are encountered in the operational setting of interest’[45]. The choice of documents to be included should result in a set that reflects the ‘the diversity of subject matter, word choice, literary styles, document formats, etc. of the operational setting for the retrieval results to be representative of the performance in the real task’[45].
2. **Task/Topic/Query:** The term ‘task’ in Information Retrieval refers to an operation on an underlying corpus. An example of such a task is ad-hoc retrieval, where the user specifies his information need through a query in order to initiate a search on a retrieval system. Since its introduction in TREC, defining a topic as an entity con-

Workshop	Document Type	Time Period	# of Docs.	# of Topics	Rel. A. Creation
NTCIR-3	Patent JPO(J)	1998-1999	697,262	31	Manual
	Abstracts(E/J)	1995-1999	ca. 1,7 million	31	Manual
NTCIR-4	Patent JPO(J), Abstracts(E)	1993-1997	1,700,000	103	Manual
NTCIR-5	Patent JPO(J), Abstracts(E)	1993-2002	3,496,252	1223	Inferred
NTCIR-6	Patent USPTO(E)	1993-2002	1,315,470	3221	Inferred

Table 1. Overview of NTCIR patent test collections (E=English, J=Japanese)

sisting of an information need and the data structure executed on a retrieval system (the query) has been widely adopted. Herein the information need statement precisely describes which documents are to be considered relevant for a specific query. Tasks are then represented through a set of topics. Since the chosen topics should resemble realistic use cases for a given task, it is often considered a best practice to involve domain experts in the design of topics.

3. **Relevance Assessments:** Given information needs and documents, the final step in the creation of a test collection consists of creating relevance assessments for each topic. Three main approaches can be distinguished for the creation of relevance assessments.

- **Manual relevance assessments:** This is a time-consuming and expensive process involving assessors, preferably domain experts, to examine a document's relevance with respect to a topic. For small collections like Cranfield, exhaustive judgments of relevance for each query and document pair can be obtained, resulting in a complete judgment. For large modern collections, it is usual for relevance to be assessed only for a subset of the documents for each query (incomplete judgment). The most commonly applied approach is referred to as 'pooling' (See [11]). The method is based on assessing relevance 'over a subset of the collection that is formed from the top k documents returned by a number of different IR systems (usually the ones to be evaluated)' [?]. In the domain of patents the manual creation of relevance assessments has been applied to tasks for cross-language retrieval [27], associative retrieval [27], and invalidity search retrieval [18].
- **Simulated relevance assessments:** The use of simulated queries and relevance assessments provides a potentially cost effective way of performing evaluation. This approach aims at replicating the actual process of retrieval based on heuristic and statistical models [41]. The primary concern lies in the realism of such approaches. The concept has been applied to a variety of tasks such as Known Item finding [10], annotation re-

trieval [23], and query generation for relevance feedback [30].

- **Inferred relevance assessments:** In this case relevance assessments are inferred from information within the corpus or information associated with the corpus. In Web IR experiments have been conducted interpreting clickthrough data (i.e. logs of a user's interaction with a retrieval system) as implicit relevance assessments [28]. In patent related retrieval, references found on patents [19, 20], and the assigned patent classes [17, 24, 25] have been utilized for relevance assessment creation.

As previously mentioned, in this paper, we propose a general methodology for the creation of topics for prior art search by inferring relevance assessments. In order to assess the feasibility of such an approach the following section will provide a brief overview of the patent system, its patenting process, and the characteristics of its documents, users, and tasks.

4 Characteristics of the patent domain

Since our proposed methodology is based on the analysis of specific aspects of the patent system, and the patent domain exhibits significant differences from other domains, the following section will provide a brief introduction to patents and outline the characteristics of the patent system.

4.1 Patent system

A patent represents a contract between a state or regional organization and the applicant to grant exclusive rights with respect to a new and useful invention (See [32] for a detailed discussion of the patent system). More specifically, a patent serves two main purposes: (1) It is a grant of the right to exclude others from making, using, offering for sale, or selling an invention in a specific country or region for a limited period of time, and (2) it discloses the described invention to the public (the informative aspect of a patent based on the doctrines of disclosure and enablement, see section B. Doctrines of Disclosure and Enablement in [34] for further information). To be granted these exclusive rights a patent has to fulfill the criteria of

patentability. In general to comply with the requirements of patentability an invention has to comply with the following four criteria.

1. **Novelty:** The criterion of novelty asks for an invention not to form part of the state of the art (i.e. it has to be 'new'). This requirement has to stand the test against all matter (patent-, non-patent literature, products, records of presentations, etc ...) made available ahead of the priority date (the date where an invention was first duly filed for protection) of a patent application.
2. **Inventive step:** The inventive step criterion requires an invention to exhibit sufficient inventive merit as opposed to merely representing a trivial extension to the state of the art.
3. **Industrial applicability:** The third requirement consists of industrial applicability of the invention.
4. **Patentable matter:** Finally the invention has to fall into the category of patentable matter (e.g. do not fall into excluded material such as literary or artistic work).

Section B IV 1/1.1 in [16] provides a more detailed coverage of these criteria in the European patent system (See [4] for the United States Patent Office (USPTO) equivalent). In order to explore how it is possible to reverse engineer the patenting process to infer relevance assessments, the next section will cover the different stages in the creation of a patent document.

4.2 Patenting process

The patenting process can be broadly divided into four main stages. Since the referencing of prior art plays a vital role in our effort we will outline the parts of the process dedicated to its identification. We will outline the patenting process based on the European patent system (see Akers [8] for a more detailed coverage of the European patent system).

1. The first stage consists of the drafting of a patent application by the applicant and the search of relevant prior art with respect to that specification. Prior to the actual filing of the patent application, this initial prior art search aims at identifying whether prior publications in patent and non-patent literature exist that might contradict with the patentability of the sought after application. Unlike the USPTO the statutes of the EPO hold 'no duty of candour' (i.e. the duty to reveal all relevant prior art to the patent office) for the applicant. As pointed out in more detail in Section 6.3 this may influence the quality of patent references [13].

2. Given that this search does not reveal any conflicting documents the second stage consists of the actual filing of the patent application with a patent office.
3. In the third stage an examiner at the patent office will conduct an examination of the application with respect to its patentability. Nowadays patent offices (as practiced by the EPO, JPO) commonly apply the concept of deferred examination. The process of deferred examination separates the lengthy procedure of substantive examination from the relatively quicker step of establishing a search report of relevant prior art. Therefore a first test of patentability under deferred examination will consist of a prior art search conducted by an examiner. All relevant prior art identified by the examiner will then form part of the search report (see figure 2 for an excerpt from such a document) for the application in question. Given that the results of this prior art search do not deny the novelty of the invention, the examination of the remaining criteria of patentability will be conducted.
4. Depending on the outcome of the examination and no withdrawal of the application from the applicant's side, in the fourth stage the patent will be either granted or denied.

4.3 Document characteristics

Patent documents are highly structured documents, which are usually broken down into a number of different sections. The exact structure of patents is defined by issuing authorities (i.e. the national and regional patent offices). The structure of patent documents therefore varies substantially between those and, due to internal revisions of that structure, also within such authorities.

In general patents can be divided into three main sections: Bibliography, disclosure, and claims. The listing below outlines these sections and their most commonly occurring subsections.

1. **Bibliographic Data:** The front page of a patent usually consists of bibliographic data such as the patent id, inventor name, applicant name, and filing date. A point of reference concerning bibliographic data is given by the WIPO standard ST.9 [5].
2. **Disclosure:** This section is aimed at providing both brief and detailed descriptions of the invention. It usually consists of the title, a subsection referring to the technological background, a summary, and a detailed description of the invention and examples of its application. By providing this

information, this part of a patent aims at fulfilling the requirements set through the doctrines of disclosure and enablement (i.e. to enable one skilled in the art to practice the claimed invention without undue experimentation).

3. **Claims:** This section forms the legally binding part of a patent application. The claims must particularly point out and distinctly claim the subject matter which the applicant regards as his invention. The reasoning is that possible infringers must be able to understand what is and is not protected.

Finally, due to their legal aspect and their nature of describing new inventions, patent documents compared to documents in other domains feature particularly different characteristics on the document, sentence, and term level.

- **Structure:** As outlined above patent documents exhibit a high level of structuring.
- **Classification:** Patent documents are classified with respect to technological aspects. The most commonly applied classification system is the International Patent Classification (IPC) [6], consisting of 70,000 classes. Other important classification systems are the European ECLA and the Japanese F-Term system.
- **Named Entities:** Induced by their functional nature, patent documents contain a high amount of named entities. Examples of explicitly denoted named entities are the inventor of the patent, specific dates relating to the patenting process, the applicant (i.e. the 'owner' of the patent), and relevant international or national patent classes.
- **Obfuscation:** For strategic reasons (e.g. to enlarge the scope of an invention and to complicate competitive analysis) the wording of patent documents may be deliberately vague or make use of general terms or obfuscating synonyms.
- **Technical and new terminology:** Due to their nature patent documents contain a high amount of technical terms. Moreover applicants commonly coin original terms to describe their inventions. Further it is noteworthy that the terminology differs significantly between above mentioned patent classes.
- **Length:** Patent documents are comparatively long documents (on average more than twenty times longer than newspaper articles [26]).

4.4 Task characteristics

As mentioned before, while patents are also informative, they are first and foremost functional documents of judicial nature. The associated tasks are therefore strongly shaped and driven by judicial and economic requirements. As a consequence of those requirements patent-related user tasks are very well defined and documented in terms of objectives, collection requirements, and course of conduction (See e.g. [14]). In contrast to other domains such as the Web, the vast majority of practitioners of patent related retrieval are professional users.

The main search tasks in patent retrieval include the following:

- **Prior art search:** The identification of prior art forms part of the Patentability (also referred to as Novelty) search type, which is probably the most frequently exercised patent search type. These searches form an essential part of the process of determining the patentability of a specific invention. In order for an invention to be viable for patenting, no prior record of a similar or identical product or process may exist. This search task aims at clarifying whether any such records exists in patent and non-patent literature that have been published prior to the filing of a patent application in question. Prior art also plays a vital role in Validity (Invalidity) searches that are exercised in order to render specific claims of a patent, or the complete patent itself invalid by identifying matter published before the filing date of the patent in question. It is of note that for this kind of searches the legally active part of the targeted patent, the claims, form the basis of the information need.
- **Freedom-To-Operate search:** This search is performed in order to assess whether a planned product or process is in danger of infringing upon someone's patent. Analogous to the Invalidity search type this task focuses solely on the claims section of a patent. Unlike the two previous tasks the searched matter is limited to granted and still active patent documents.
- **Competitive analysis:** This search type aims at identifying financial, organizational, or technological information based on an analysis of a competitor's patents.

Characteristic of the above listed patent related tasks are the following points:

- **Relevance of documents is determined by topical and judicial aspects:** Often the underlying judicial requirements will determine the relevance of a document with respect to a search task. For a certain search task, a patent document's content

might be highly relevant to the search but the document itself can still be not relevant since it is no longer in force (e.g. due to missed renewal payments, expiration of the patent). Relevance for most patent related tasks is therefore a notion of topical as well as judicial relevance.

- The high value associated with patents (e.g. potential damages of infringement of several 100 millions of Euros) leads to a high importance on recall within patent searches. Ideally most patent search tasks aim at achieving 100 % recall, since missing a single relevant document could result in later litigation
- Potential legal responsibility requires many patent related tasks to be process traceable. Items such as the exact properties of the underlying corpus and the functioning and implementation of the retrieval method have to be known to a searcher and kept record of.

Concluding it can be said that these specific characteristics of patent related user tasks represent some of the major challenges for performing Patent Information Retrieval.

5 Methodology

This section introduces our methodology for creating a Prior Art test collection. The following high level steps form the basis of the creation of topics for prior art search. Given a set of patent documents from a given time period: a corpus, topics and corresponding relevance assessments can be created as follows:

1. Define the set of documents forming the corpus.
2. Define the pool of documents, outwith the defined time period of the corpus that form potential information needs.
3. Select a patent document from this pool.
4. Extract the set of references, which refer to prior art for this document.
5. Identify the references which exist in the corpus.
6. For those references that exist, mark these documents as relevant documents for the prior art search.
7. Based on these references define a topic with the patent application (without references) or a subset of its text as query, the extracted references as relevance assessments and the definition of relevant prior art (e.g. 'Guidelines for Examination in the EPO B X 9.2' [16] for European patents, [4] for US patents) as information need.

8. Repeat steps (3) to (7) for each of the documents in the pool defined in step (2).

In the following we will outline specific considerations of these steps for the corpus, tasks, and relevance assessments.

5.1 Corpus

Step (1) of our methodology consists of defining a set of documents to form the corpus. As stated in section 3 such a corpus should be representative of the operational setting of interest. In this case the resulting document set should resemble a realistic source of research for an Intellectual Property practitioner conducting prior art search. The judicial basis of prior art search defines all patent and non patent literature published prior to the filing date of a patent as potentially relevant documents. A valid corpus would therefore consist of a subset of these documents. Concerning the criteria of representativeness such a set should resemble the data sources utilized by practitioners of prior art search. Commonly researched information sources for patent search are represented by free databases such esp@cenet¹, WIPO², and the USPTO database³ as well as subscription databases such as Thomson's Derwent World Patents Index⁴. Potential sources for non-patent literature are periodicals, repositories of scientific publications such as Medline⁵ and the IP.com prior art database⁶. A chosen corpus therefore preferably should resemble one of these sources in order to allow practitioners and researchers of the patent domain to easily relate to research conducted on the test collection.

5.2 Relevance assessments

This section relates to steps (2) to (6) of our methodology. The idea of interpreting references as relevance assessments for prior art search as presented in this work is based on a proposal from IP practitioners at the 2007 Information Retrieval Facility Symposium. The focus of this section lies in a brief exploration of the concept of patent references and an analysis of the justification of inferring relevance assessments from them.

Patents contain citations to other prior published patent and non-patent literature. The functional aspect lies in referring to the most relevant prior art upon which the patent builds. In the case of patent documents the placement of a citation first of all denotes that the concepts described in both documents

¹<http://ep.espacenet.com/>

²<http://www.wipo.int/pctdb/en/>

³<http://patft.uspto.gov/>

⁴<http://scientific.thomson.com/products/dwpi/>

⁵<http://medline.cos.com/>

⁶<https://priorart.ip.com/>

are semantically strongly related to each other, and secondly that the concepts described in the cited document logically pre-date those mentioned in the citing document. The primary motivation for patent citations is the expression of this relationship. This stands in contrast to web-based document links that can be of semantic or navigational nature, with the underlying motivation ranging from the expression of relevance to the direction of traffic due to commercial interests. Detailed information concerning the interpretation of the citation process and resulting references for EPO data has been provided by Akers [9]

The justification for reverse engineering relevance assessments from the references within a patent document is based on the following:

- The patent references found on patent documents issued by a patent office are set by its patent examiners. The subject and legal expertise of the examiner at the patent office allows for qualified assessment of relevance from his or her side with respect to the prior art search task.
- The legal specification setting the criteria for valid reference matter can be interpreted as a definition of relevance for an information need (e.g. European Patent Convention [44] Rule 44, Article 92(1), and Article 54).
- Additional guidance concerning the interpretation of references is provided by examination manuals (e.g. USPTO Manual of Patent Examining Procedure [43]) provides a further precise description of the nature of the stated form of relevance. This is exemplary demonstrated through an excerpt:
All documents cited in the search report are identified by placing a particular letter in the first column of the citation sheets. ... Where a document cited in the European search report is particularly relevant, it should be indicated by the letter 'X' or 'Y'(, Guidelines for Examination in the EPO B X 9.2' [16])

Upon successful extraction of the patent references the final stage of our methodology consists of defining the topics.

5.3 Topic

On successful completion of steps (2) to (6), step (7) of our methodology consists of the definition of the topics for our prior art search task. As mentioned before the identification of valid prior art forms part of several patent related search tasks. We will recapitulate on the two most common tasks focused on the search of prior art below.

1. A 'novelty search' or 'patentability search' is a prior art search that is conducted by patent attorneys, patent agents or patent examiners in the process of a patent application filing. This type of search aims at determining if the invention is novel. In the course of this search the total of the claims of an application and the disclosure will form part of the information need.
2. A 'validity search' or 'invalidity search' is a search for prior art based on a patent that has been granted. The purpose of a validity search is to try to identify prior art that the patent examiner overlooked in order to render a specific claim or the complete patent invalid. This might be done by an entity infringing, or potentially infringing, the patent. In the novelty search type the information need is usually centered around one or more claims and not the complete patent application.

While both described search types are potentially viable for our methodology our initial focus will be placed on the patentability search task, as the source of our relevance assessments (i.e. patent references placed by an examiner) presents the direct result of such a search.

A topic for patentability search shall then be defined as consisting of the following three parts:

1. **Statement of the information need:** The statement of the information need will be based on the written guidance (i.e. a description of what documents qualify as prior art) on which the examiner's original search was based (See EPO B X 9.2' [16] for European patents), and the appropriate legal definition of potential prior art (European Patent Convention [44] (Rule 44, article 92(1), and article 54).
2. **Query:** A query will consist of the complete or a subset of the text comprising a patent application
3. **Relevance assessments:** The extracted patent references forming the relevance assessments for the application.

When using the complete text of an application the task will represent a document to document retrieval task. In the case that the patent reference outlines (1) in respect to which claim a reference is made, and (2) what part of the referenced document is relevant, this task could be refined to a claim to document, or claim to passage retrieval task.

In the following section we will explore the applicability of our methodology on a corpus of European patent documents.

6 Towards Applying the Methodology to European Patent Documents

This section explores the viability of a patent data source as the basis for our methodology. The set of documents to be explored consists of documents issued by the European Patent Office documents (EP documents) in the time period from 1978 until March 2008.

6.1 Corpus

The set of these EP documents, resembling those hosted by esp@cenet, represents one of the most important data sources in patent search. In the remainder of this section we will briefly explore some key characteristics of this dataset. The collection consists of 3.6 million documents representing 1,896,483 patents and patent applications. During the patenting process a patent application will depending on its status be published several times under different patent IDs. In the European patent system a patent publication is published with a numerical identifier and an extension, the kind code that refers to the status or type of the publication. The listing below explains the most common kind codes used by the EPO.

- **A1:** Publ. of an application with a search report.
- **A2:** Publ. of an application without search report.
- **A3:** Publ. of a search report.
- **A4:** Publ. of a supplementary search report
- **B1:** Publ. of a granted patent.
- **B2:** Publ. of a patent after modification.

A granted patent published as B1 document therefore will have been prior published as an A1 or A2 document. Such documents sharing the same numerical identifier, but a different kind code are referred to as belonging to the same 'Patent Family'. Due to this our dataset consists of 3.6 million documents but represents only 1.9 million patents and patent applications. Table number 2 shows the frequencies of these kind codes in the collection. It is of note that for the prior art search task granted patents as well as patent applications are potentially relevant. Moreover, even though A3 and A4 designated documents consist only of a listing of patent references, these documents should also be included into the collection as they form part of established patent information sources such as esp@cenet. EP patent documents may occur in one of the three official languages of the EPO: English, German, or French. Table 3 outlines the occurrence of these languages in our data set. Another important aspect of the underlying data is given by the fact that

Kind code	# of documents
A1	1226849
A2	678434
A3	686075
A4	157957
B1	890436
B2	13286
Other	10032

Table 2. Frequency of most common kind codes

Total number	ENG	GER	FR
3,631,954	2,549,633	848,471	232,950

Table 3. Distribution of languages among patent documents

the complete text of the patent document is only available for a part of these 3.6 million documents. The availability of text for segments of the patent is given in the Table 4. We will limit the documents to be in-

Total	Abstract	Description	Claims
3,631,954	2,116,081	1,075,162	2,189,941

Table 4. Availability of text for patent documents

cluded in our collection to those containing at least the full text of the claims section; therefore the maximum number of documents to be included in the collection is 2,189,941.

6.2 Task and topics

In this section we will explore potential tasks and topics for patentability prior art search based on the underlying corpus.

Following our methodology a statement of the information need will be based on the written guidance on which the examiner's original search was based (See EPO B X 9.2' [16] for European patents), and the appropriate legal definition of potential prior art (European Patent Convention [44] (Rule 44, Article 92(1), and Article 54). A query will consist of a complete patent application or part of a patent application. Based on this the underlying corpus lends itself to define tasks considerable of the following characteristics:

- **Language:** The existence of three different languages is one property that could be used to define sets of language specific tasks and topics (e.g. for cross-language information retrieval).

- **Classification:** Since patent documents filed under different patent classification codes differ significantly in terms of terminology, the importance of graphics, and the technological interpretation of relevance it seems appropriate to define tasks with respect of these classes. Such an approach would enable to explore the effectiveness of retrieval models across different classes or to define specific tasks for classes with specific properties (e.g. occurrence of Markush structures [39])
- **Reference categories:** The categorization of references according to their origin and level of relevance allows the definition of tasks with respect to this. Such tasks could form the basis to explore the effectiveness of retrieval models for state of the art ('A') denoted prior art and highly relevant ('X','Y') denoted prior art.
- **Document granularity:** Since references made in a European search report (see Figure 1) are marking the relevant sections of the referenced document with respect to the relevant claims, tasks on this test collection could be defined as document to document, claims to document, or claims to passage retrieval tasks.

The limiting point concerning the definition of those tasks and topics lies in the number of available relevance assessments. Therefore the distribution of these relevance assessments will be explored in the following section.

6.3 Relevance Assessments

EPO references are considered to be of high quality, since there is no duty of candour for the applicant, and therefore all references found on European patents are solely based on an examiner's judgment. It has been found [13] that patent references supplied by the applying party are generally of lower relevance than references identified by an examiner. This might be based on the fact that an applying party might be less motivated of supplying highly relevant prior art citations. The use of European references therefore seems to be very promising with respect to the inference of relevance assessments.

In the underlying data set a total of 5,096,362 references, representing 3,063,246 documents can be found on EP documents. Table 5 on page 10 shows the distribution of referenced patent documents among issuing authorities.

As can be seen from Table 5, 491,251 of all cited documents are EP. These documents form the pool of relevant documents for our test collection. In the underlying dataset 1,106,362 references to these documents are found.

References in the EPO patent system are categorized with respect to their level of relevance, origin, and type. The listing below explains these categories:

- **'X':** Particularly relevant documents.
- **'Y':** Documents that are particularly relevant in combination with another document.
- **'A':** A document cited that represents the state of the art not prejudicial to the novelty or inventive step of the claimed invention.
- **'P':** Documents published on dates falling between the date of filing of the application being examined and the date of priority claimed, or the earliest priority if there is more than one.
- **'E':** Any patent document bearing a filing or priority date earlier than the filing date of the application searched.
- **'D':** Documents cited in the text of the application (i.e. references included by the applicant).

Up to three of these categories may be assigned to a patent reference. The distribution of these is provided in Table 6. As pointed out before, these categories

Categ.	/	P	D	PD	Total
X	219,610	24,795	11,722	1,085	257,212
A	568,412	18,731	64,732	1,142	653,017
Y	140,705	4,402	16,010	435	161,552
E	15,667	/	/	/	15,667

Table 6. Distribution of references among categories

could be utilized for the formulation of tasks focused on different 'levels' of prior art relevance. Finally concerning the viability of formulating tasks and topics based on the set of documents we will explore the distribution of the frequency of these citations within the corpus. On average there are 0.58 EP references per patent application and granted patent. Figure 2 shows the distribution of these references among all documents.

These frequencies are also presented in Table 7 on page 10. As can be seen from the table there are 43,306 documents bearing four or more references. These documents could form the basis for 43,306 topics with at least four assessments of relevance.

The distribution of 'EP' references suggests that many patents exhibit few references which can be used as inferred relevance assessments. As discussed by [35] the usage of topics created with a low number of relevance judgments can be problematic and is a point to be considered in our future work.


 European Patent Office		EUROPEAN SEARCH REPORT	Application Number EP 05 01 5582
DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
X	EP 0 428 322 A (THE GENERAL ELECTRIC COMPANY, P.L.C) 22 May 1991 (1991-05-22) * the whole document *	1-22	H04L12/56 G08C17/02
D,A	DE 195 02 839 C1 (BRENDDEL, WOLFGANG, DIPL.-ING., 74564 CRAILSHEIM, DE) 5 June 1996 (1996-06-05) * column 3, line 66 - column 6, line 11 * -----	1,14,21	

Figure 1. Excerpt from a European search report

US	DE	EPO	FR	GB	WO	CH	BE	AT	NL
1,137,213	530,279	491,251	265,693	235,762	215,809	88,584	35,149	12,271	10,084

Table 5. Frequency of references on EP documents: Top ten countries

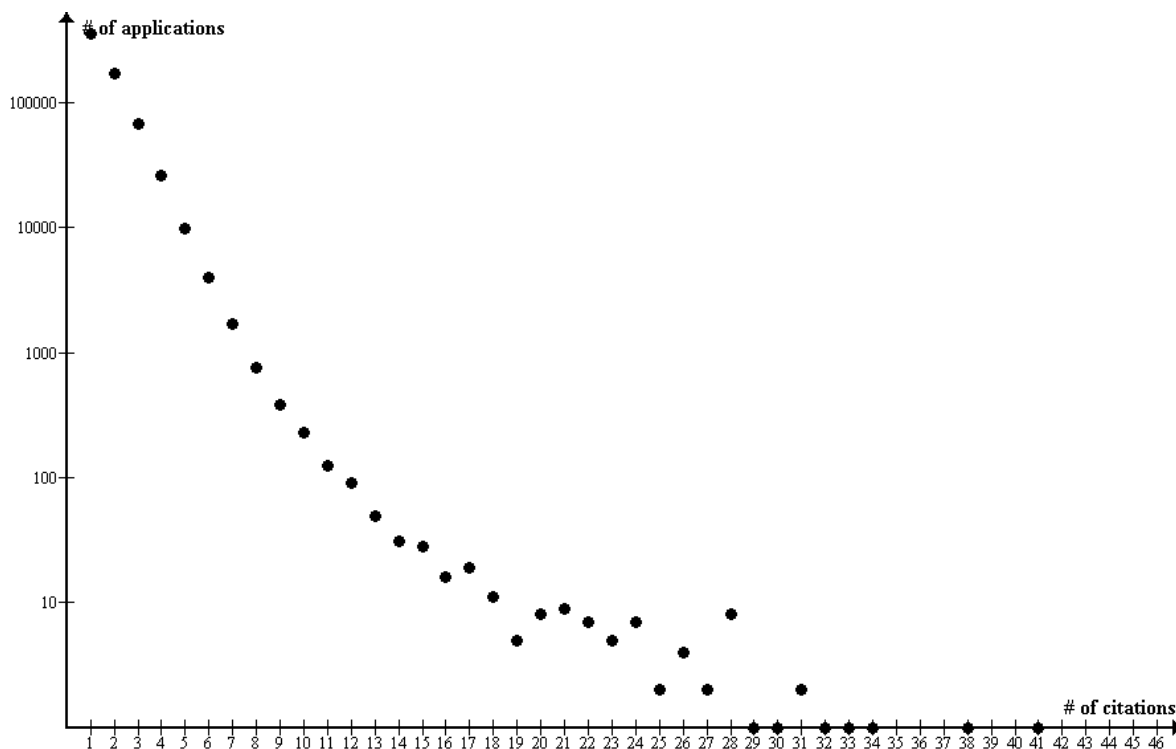


Figure 2. Frequency of EP citations

# of cit.	1	2	> 2	> 3	> 4	> 5	> 6	> 7	> 8	> 9	> 10
# of docs.	357,387	168,896	111,397	43,306	17,279	7,430	3,486	1,798	1,043	664	435

Table 7. Distribution of EP references

7 Discussion

Our initial exploration of the characteristics of patents shows that references found on those documents, especially so on European patents, can be interpreted as an explicit statement of relevance.

An analysis of references in the corpus of EP documents indicates that their utilization as means of cost effective creation of realistic relevance assessments for the prior art search task seems promising. As mentioned before the low number of relevant documents per topic for many patent documents may be problematic. Exploring the impact of this and anticipating ways of mitigating it will form part of our future work.

Additionally our planned work will focus on exploring and refining the prior art search task with respect to the options outlined in section 6.2. This will allow the assessment of the strengths and weaknesses of different retrieval methods on a variety of tasks and with respect to the particularities of technical domains (e.g. the importance of Markush structures for chemical patents, the varying importance of pictures, particularities of patent classes, etc ...).

The proposed methodology lends itself to be applied to patent corpora issued by other patent offices such as the United States Patent Office (USPTO) and the State Intellectual Property Office of the People's Republic of China (SIPO). Exploring the applicability of our methodology on such data sources will form another directive in our future work.

Finally as a validation step aimed to uncover potential weaknesses and strength of our methodology it will be vital to conduct a series of initial retrieval experiments on a pilot test collection. Concerning this a number of challenges remain to be considered. First as noted before the format of patent documents varies significantly not only between issuing patent offices but also on a lower degree within patents issued by the same office. This poses a major obstacle concerning the utilization of the structure and bibliographic data of patent documents. Even when focusing on documents issued by a single patent office, the evolution of the format of patent specifications and associated standards renders this task difficult, and will require extensive research concerning this aspect. Second the question of valid evaluation measures remains to be discussed. The high importance of recall and the low number of relevant documents per topics form interesting aspects of this challenge. The approach described could be employed at any one of the major forums in order to provide a track in patent retrieval. While as pointed out a number issues remain, it is anticipated that these can be resolved satisfactorily to enable the creation of reliable and high quality prior art patent test collections.

8 Acknowledgements

The authors would like to thank Keith van Rijsbergen, Matrixware Information Services⁷, the Information Retrieval Facility⁸ (IRF), and Fairview Research⁹ for their support of this work. We also would like to specifically thank Henk Tomas for the insightful discussions concerning the interpretation of references and David Santamauro for his support concerning the interpretation of the patent data format.

References

- [1] The cross-language evaluation forum (clef). <http://www.clef-campaign.org/>.
- [2] National institute of informatics test collection for ir systems (ntcir). <http://research.nii.ac.jp/ntcir/>.
- [3] Text retrieval conference (trec). <http://trec.nist.gov/>.
- [4] United states code of law, title 35, §102-105. http://www.uspto.gov/web/offices/pac/mpep/document/s/appx1_35_U.S.C.102.htmusc35s102.
- [5] *Handbook on Industrial Property Information and Documentation*, chapter 3, page 3.9.0 to 3.9.12. World Intellectual Property Organization (WIPO), 2008.
- [6] S. Adams. Using the international patent classification in an online environment. *World Patent Information*, 22(4):291–300, Dec. 2000.
- [7] P. Aghion, N. Bloom, R. Blundell, R. Griffith, and P. Howitt. Competition and innovation: An inverted-u relationship. *The Quarterly Journal of Economics*, 120(2):701–728, May 2005.
- [8] N. J. Akers. The european patent system: an introduction for patent searchers. *World Patent Information*, 21(3):135–163, Sept. 1999.
- [9] N. J. Akers. The referencing of prior art documents in european patents and applications. *World Patent Information*, 22(4):309–315, Dec. 2000.
- [10] L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 455–462, New York, NY, USA, 2007. ACM.
- [11] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 619–620, New York, NY, USA, 2006. ACM.
- [12] C. Cleverdon. *Readings in information retrieval*, chapter The Cranfield tests on index language devices, pages 47–59. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.
- [13] P. Criscuolo and B. Verspagen. Does it matter where patent citations come from? inventor versus examiner citations in european patents. Research Memoranda 017, Maastricht : MERIT, Maastricht Economic Research Institute on Innovation and Technology, 2005.

⁷<http://www.matrixware.com>

⁸<http://www.ir-facility.org/>

⁹<http://www.fairviewresearch.com/>

- [14] M. R. David Hunt, Long Nguyen. *Patent Searching: Tools & Techniques*. Wiley, Hoboken, February 2007.
- [15] J. Eaton and S. Kortum. Trade in ideas patenting and productivity in the oecd. *Journal of International Economics*, 40(3-4):251–278, May 1996.
- [16] European Patent Office (EPO). *Guidelines for Examination in the European Patent Office*, December 2007.
- [17] C. J. Fall, A. Töröcsvári, K. Benzineb, and G. Karetka. Automated categorization in the international patent classification. *SIGIR Forum*, 37(1):10–25, 2003.
- [18] A. Fujii, M. Iwayama, and N. Kando. Overview of patent retrieval task at ntcir-4. In *Proceedings of NTCIR-4 Workshop Meeting*, 2004.
- [19] A. Fujii, M. Iwayama, and N. Kando. Overview of patent retrieval task at ntcir-5. In *Proceedings of NTCIR-5 Workshop Meeting*, 2005.
- [20] A. Fujii, M. Iwayama, and N. Kando. Overview of the patent retrieval task at the ntcir-6 workshop. In *Proceedings of NTCIR-6 Workshop Meeting*, pages 359–365, 2007.
- [21] Gambardella, Alfonso, Harhoff, Dietmar, Verspagen, and Bart. The value of european patents. *European Management Review*, 5(2):69–84, 2008.
- [22] K. Idris. Intellectual property: A power tool for economic growth. Technical Report Publication N0 888, ISBN 92-805-1113-0, WIPO, Geneva, 2003.
- [23] M. Inoue and N. Ueda. Retrieving lightly annotated images using image similarities. In *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, pages 1031–1037, New York, NY, USA, 2005. ACM.
- [24] M. Iwayama, A. Fujii, and N. Kando. Overview of classification subtask at ntcir-5 patent retrieval task. In *Proceedings of NTCIR-5 Workshop Meeting*, 2005.
- [25] M. Iwayama, A. Fujii, and N. Kando. Overview of classification subtask at ntcir-6 patent retrieval task. In *Proceedings of NTCIR-6 Workshop Meeting*, pages 366–372, 2007.
- [26] M. Iwayama, A. Fujii, N. Kando, and Y. Marukawa. An empirical study on retrieval models for different document genres: patents and newspaper articles. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 251–258, New York, NY, USA, 2003. ACM.
- [27] M. Iwayama, A. Fujii, N. Kando, and A. Takano. Overview of patent retrieval task at ntcir-3. In *Proceedings of NTCIR-3 Workshop Meeting*, 2002.
- [28] T. Joachims. Evaluating retrieval performance using clickthrough data. In *In Proceedings of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*, pages 79–96, 2002.
- [29] K. S. Jones. *Information Retrieval Experiment*, chapter 13, pages 256–284. Butterworths, London, 1981.
- [30] C. Jordan, C. Watters, and Q. Gao. Using controlled query generation to evaluate blind relevance feedback algorithms. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 286–295, New York, NY, USA, 2006. ACM.
- [31] N. Kando and M.-K. Leong. Workshop on patent retrieval sigir 2000 workshop report. *SIGIR Forum*, 34(1):28–30, 2000.
- [32] E. W. Kitch. The nature and function of the patent system. *Journal of Law & Economics*, 20(2):265–90, October 1977.
- [33] J. Lerner. The importance of patent scope: An empirical analysis. *RAND Journal of Economics*, 25(2):319–333, Summer 1994.
- [34] R. P. Merges and R. R. Nelson. On the complex economics of patent scope. *Columbia Law Review*, 90(4):839–916, 1990.
- [35] A. Ritchie, S. Teufel, and S. Robertson. Creating a test collection for citation-based ir experiments. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 391–398, Morristown, NJ, USA, 2006. Association for Computational Linguistics.
- [36] K. G. Rivette and D. Kline. *Rembrandts in the attic: unlocking the hidden value of patents*. Harvard Business School Press, Boston, MA, USA, 2000.
- [37] S. Robertson. On the history of evaluation in ir. *Journal of Information Science*, 34(4):439–456, 2008.
- [38] T. Saracevic. Evaluation of evaluation in information retrieval. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 138–146, New York, NY, USA, 1995. ACM.
- [39] E. S. Simmons. Markush structure searching over the years. *World Patent Information*, 25(3):195–202, Sept. 2003.
- [40] H. Small. Tracking and predicting growth areas in science. *Scientometrics*, 68(3):595–610, Dec. 2006.
- [41] J. Tague, M. Nelson, and H. Wu. Problems in the simulation of bibliographic retrieval systems. In *SIGIR '80: Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, pages 236–255, Kent, UK, UK, 1981. Butterworth & Co.
- [42] J. Tait. Information retrieval facility symposium in vienna. *SIGIR Forum*, 42(1):67–67, 2008.
- [43] United States Patent Office (USPTO). *Manual of Patent Examining Procedure (MPEP)*, July 2008.
- [44] D. Visser. *The annotated European Patent Convention. (10th revised ed.)*. H. Tel Publisher, Veldhoven, The Netherlands, 2003.
- [45] E. M. Voorhees. The philosophy of information retrieval evaluation. In *CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, London, UK, 2002. Springer-Verlag.
- [46] E. M. Voorhees. *TREC: Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing. MIT Press, September 2005.
- [47] E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 316–323, New York, NY, USA, 2002. ACM.
- [48] J. Youtie, M. Iacopetta, and S. Graham. Assessing the nature of nanotechnology: can we uncover an emerging general purpose technology? *The Journal of Technology Transfer*, 33(3):315–329, June 2008.