Journal of Intellectual Property Rights
Vol 29, January 2024, pp 50-56
DOI: 10.56042/jipr.v29i1.1435

A Recapitulation of Patent Retrieval

K S A R Mohamed Ali† and K Kathiresan

Department of Pharmacy, Annamalai University, Annamalainagar – 608 002, Tamil Nadu, India

Received: 26th January 2023; revised: 11th May 2023

The search for patentability is a critical element in the patent process, and failure to find a certain relevant patent may result in costly legal consequences. This paper aims to review the studies that have been conducted in result to improve the significance efficacy of patent information retrieval, which is focused on creating techniques and methods for retrieving patent documents in response to a specific user query. The prior-art search is a crucial step in the patent retrieval process, the choice of pertinent search terms is essential to the task's success. In this review, we develop prior-art queries derived from query patents using query expansion to pseudo relevance feedback in result of increase the retrievability of patents. The review of this literature lists numerous studies that have been done to enhance current information retrieval methods or use conventional methods at different phases of the patent retrieval task in order to achieve better results in the task. This research focuses solely on the literature on patent text and image retrieval. Given the different methods and frameworks available, as well as their limitations, there is a lot of room for further research in the field of patent retrieval techniques.

Keywords: Patent Retrieval, Patent Images, Queries, Prior Art Search, IPC, Boolean Search

By giving innovators a special monopolistic hold over the market value of their creations, the Intellectual Property (IP) method encourages the dissemination of innovative concepts and technologies. Patents have become an essential resource for any innovative company, and as global competition has increased, companies have begun to align their business strategies with their intellectual property strategies.[1] The need for efficient systems to handle such lots of data will be unavoidable as the percentage of filed patent applications rises year after year. Almost every patent analysis tasks rely on Patent Retrieval (PR). A subcategory of information retrieval called PR focuses on creating strategies and methods for locating patent data in response to a specific search query with the development of information and communication innovation, patent searching has shifted from a manual catalogue-based system to an online system.

Patent Retrieval

The tasks performed by patent data users can be classified into patent retrieval, analysis and monitoring. The main goal in Patent Retrieval is to find all the prior arts reliable to a given patent application. Depending on the end result, there are numerous names that can be used to describe the patent retrieval task, such as novelty search, infringement search, invalidity search, patentability search, freedom-to-operate search, due diligence search, and so on.[1]

According to the nature of the search task, the goals, relevance judgments, and effectiveness requirements vary greatly. Before applying for patent application, the idea should be refined by various searches like technology survey, researchers, information professionals, pre filing patentability search and so on (Fig. 1). If the search scoring function is effective then only it approves the idea and go to further process, these searches are carried out at various life cycle stages of the patent documents.

Researchers are motivated to create methods and approach for effective and efficient patent data retrieval by the increasing number of patent-associated data and the ever-increasing need to obtain this information by various types of users. These users can be patent professionals, academic research communities, industrial venture capitalist, managers, patent attorneys, and investors, etc. There are various research areas in patent retrieval and mining such as automated patent classification, evaluation of patent retrieval, image-based patent retrieval, patent text retrieval, multilingual patent retrieval and classification, etc. There is still a disconnect between methods widely used in patent retrieval and web search engine research.

†Corresponding author: Email: mohamed.ali21574@gmail.com

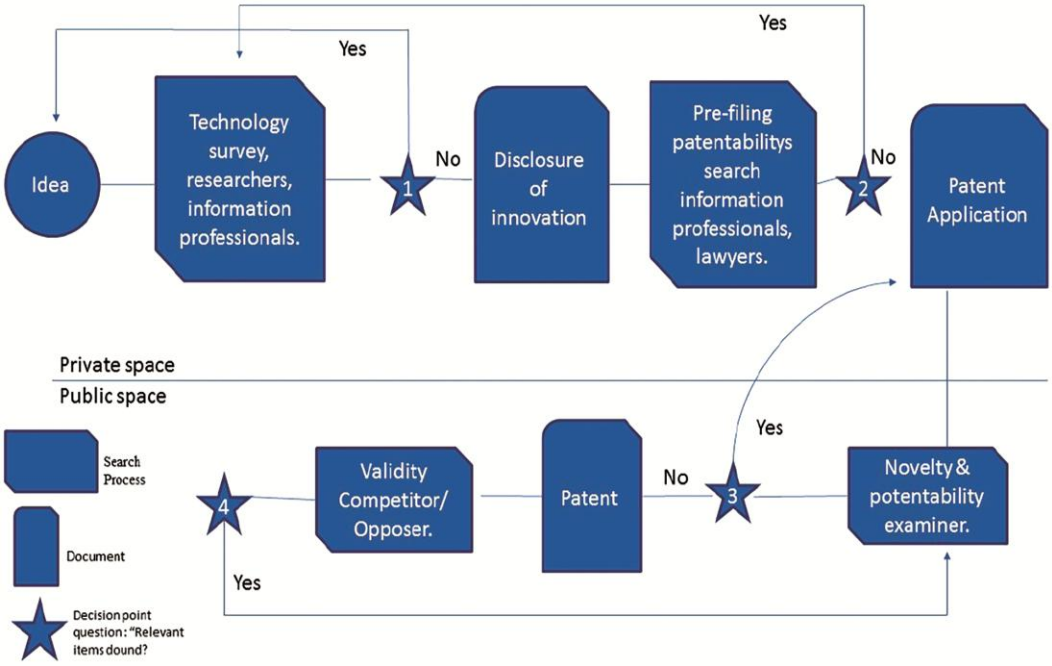

Fig. 1 — Process of patent retrieval

Information Retrieval

Information retrieval is the study of the analysis, structure, storage, organisation, searching, and retrieval of information. The primary goal of an information retrieval model is to locate relevant knowledgebased information or documents that meet the needs of the user. Recall and precision are two traditional measures of effectiveness. Precision indicates the percentage of pertinent documents among the ones that were returned, whereas recall shows the percentage of pertinent documents among the returned documents.[1] The goal of this literature review is to comprehend the current research being conducted to improve the productiveness of patent text information retrieval. It is assumed that an IR system will retrieve documents from a corpus. We believe that some IR system user has a data need that he conveys as a query. The user is informed that at least 0 documents were found, indicating that they may be relevant, and it is concluded to submit this to the IR system being investigated. IR does not cover the actual task of retrieving or duplicating the document.[2] Precision indicates the percentage of pertinent documents among the ones that were returned. Recall shows the percentage of pertinent documents among the returned documents. However, based on the nature of the search area, relevance in the patent domain contains different synonyms,

Novelty Search

If a document includes any details about prior art associated with the invention, it is relevant.

Search for Validity/Invalidity

A document is appropriate if it includes any data which could render one or additional of the patent's claims invalid.

Freedom to Operate

A manuscript is appropriate if it includes any claims that would constrain or restrict the designed operations. It is crucial to note that a patent searcher would not only want to find pertinent documents, in contrast to how most IR system users would view such an information need[4]. If none were discovered, it would also be desirable to establish with some certainty that relevant documents don't exist. Indeed, finding just one applicable 'kill' data record may be enough to put an end to the search. The scoring function having the database of patent information retrieved from the corpus index, where the query was entered by an innovator, the scoring function correlate the query with the corpus database and shows the Score value (Fig. 2).

Patent users have one intention in mind as they get ready to delve into a sizable patent data compilation: to identify the data that are most pertinent to their research. In order to do this, search queries with a

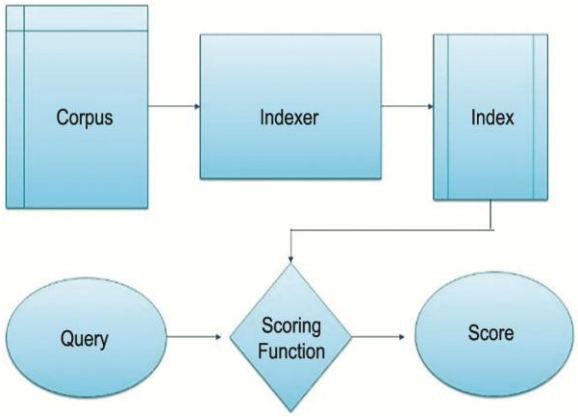

Fig. 2 — Basic architecture of an IR system

string of search terms and Boolean operators were created, which produced documents that were pertinent. A fundamental limitation of keyword ("lexical") matching for document retrieval is the process's inability to identify documents that do not include those keywords for a variety of reasons. By a Boolean search engine, less can be done to solve this issue among the employing experts in the field who make an effort to formulate questions using all possible synonyms that may be utilized to describe an invention.

Patent retrieval is extremely complex and problematic for researchers. Patent retrieval demands effective and efficient techniques as it is a multi-topical task and performed on diverse and large dataset a complete patent document is used as a query to represent the information need. Where a keyword search may not produce the desired results Patent retrieval is a recall task because missing even a single relevant document while conducting a patentability or freedom to operate search can result in severe financial consequences due to a patent infringement lawsuit. Many times retrieval techniques are only restricted to patent classification codes (IPC). This traditional approach is too general to meet the needs of users.

Boolean Search

From the literature, it is evident that Boolean search has certain limitations. To overcome the similarity issue between queries and patent documents caused by Boolean search, Vector Space Model (VSM) is also suggested where terms, queries and documents are represented as vectors. Gerard Salton proposed this model first. The vectors' dimension relates to the unique terms in the document collection, and a value represents the frequency of this term. In the search space, a user query is also regarded a document[1]. The query vector is compared to the document vector to determine similarity using the cosine of the angle between both vectors. In most studies, the entire patent application was considered in order to retrieve relevant patents for prior art. However, a novel approach is suggested which looks at the prior art from the inventor's perspective and consider ideas (partial application) to form a query rather than full application.

Limitations in the Patent Domain

Patents usually go into publication 18 months after their initial filing date, so they are not accessible to the public during that time. Patent offices manually enter data, so mistakes can happen like incorrect indexing, incorrect classification, and misspelt words. The technical meanings of words used in patents are distinct from their common usage. A patent application may not be categorised by patent examiners into the best categories. A company might have changed its name or needed patents through acquisition or licencing.

Document Type

The majority of patent datas are text-based. They are very well organised and have common components, such as the invention's title, abstract, history, claims and description. That description section describes the invention's technical specifications and possible embodiments in detail. Most important part is claim section, because it represents the scope of the inventor's protective security and thus encodes the true value of the patent.[3] They also include a variety of data types (such as images, text, flow charts, and formulae) as well as a wealth of bibliographic information and metadata (such as citations, examiners, assignees, inventors, classification codes, filing/publication dates, and addresses).

Patent Retrieval Process / Aim/ Search

When a search request is made, PR's objective is to return pertinent patent documents (query).

(i) This request may be in the form of a memo, a list of keywords, or an entire text document.
(ii) In order to get related work or refute a portion of its claims, receive associated patents to a particular patent application.
(iii) Investigate patent filing action in a particular technology.
(iv) Investigate a given company's competitive landscape by searching of filing patent at other

companies comparable to the patent of given company.

The NTCIR and CLEF-IP statistics are focused on prior art search. This task is critical because that filed patent is required in all patent offices which has to be non-obvious, novel and non-abstract ideas. Prior art searches are carried out at various points in the patent document life-cycle, by various stakeholders, for a variety of situations, and for a limited time.

Related Work Search

In order to find all prior art that is pertinent to the invention, inventors and prosecutors search for it during the pre-grant stage. Furthermore, when filing a new application, some patent offices require inventors to provide a disclosure statement submitted by an applicant listing all relevant publications.

Patentability Search

Patent examiners conduct patentability searches during the examination stage to ensure that the suggested ideas are non-obvious, novel and non-abstract. A search report containing all pertinent publications located would be the result of this task. There will be a special code next to each entry in this report designating whether it is an associated publication or even a novel publication. Additionally, examiners would point out which verses or facts from the retrieved journals are pertinent. The patent office has the authority to approve, deny, or ask the applicant to amend their application. based on the results of the search. Patent prosecutors conduct a patentability search as a sanity check.

Infringement Search

In order to ascertain whether a proposed or existing product violates any other already published patent claims, product clearance searches are conducted (s). Patent owners need this kind of search to determine whether a third party does have product with attributes that meet the requirements of at least one of their patent applications. If this is the case, they may negotiateor sue a licence with the infringing party.

Freedom to Operate Search

This public relations task goes apart from infringement detection. Investors and R&D managers must ensure that a patent or piece of IP is not violated by the proposed products, but also to make sure they are free to patent these goods without being concerned that earlier prior art might render such inventions ineligible.

Invalidity Search

Patents give their stakeholders monopoly rights over the commercial benefit of granted inventions, businesses and other parties frequently keep an eye on patents granted to rival firms or to technologies that are relevant to their own to maintain competitive advantage. As a result, an invalidity search is carried out in order to locate published works that Publications that the patent office failed to find during the search for patentability.

Technology Survey

Another PR task involves business managers asking search specialists to compile a patent documents survey based on the memo they've prepared from somewhere. (example: article, magazine, news). The basic scenario in the NTCIR-3 PR task was limited to patent literature, with the assumption that A gathering of technical papers makes up a patent document.

Patent Image Retrieval

Images and patent data are crucial for describing an invention's novelty (Fig. 3). However, the image

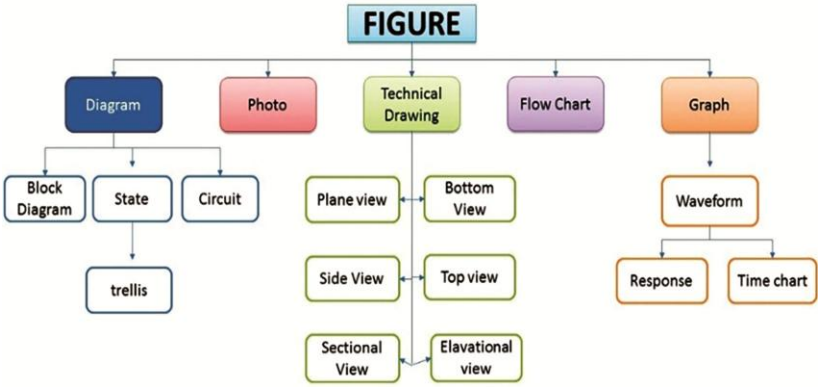

Fig. 3 — The patent technical drawing

retrieval capabilities of current patent collections make it nearly impossible to conduct searches on patent images. Despite the fact that the patent documentation contains both text and graphics/images that describe the patent, the search ignores the graphics/images information. As a result, it is typical for any given patent search to yield a large amount of information, much of which is either irrelevant or useless to the user. Adequate time and effort are needed for a user to search for patent information, which inevitably hinders their ability to do their jobs more effectively.[4] According to research, "graphic representations or diagrams can occasionally be worth more than ten thousand words." This reflects the graphical significance information, and it also indicates how crucial graphics/image retrieval is typically, binary images are used as figures in patent documentation. No colour or texture can be found on them. Using a query image as a starting point, you can search a database of images for related patent images, Patent Image Retrieval (PIR) techniques are used. The methods assist both users and experts in the field of patents in comprehending the information contained in patent images and in locating related patent images. In a classic patent image retrieval procedure, a user submits a patent pictures as an image query in order to retrieve related images in the patent directory.

To find any matching images, a patent image search is typically conducted manually. The rate at which the correct information is discovered using this manual method is low. Since Content-Based Patent Image Retrieval (CBPIR) is a more effective way than image processing-based techniques to get around this restriction, it is employed to recognise and find matches in patent images. Image matching methods are widely used for CBIPR systems, according to the literature (Fig. 4).[4] Due to their ability to match images with invariant features, some methods known as affine transformation algorithms are frequently used. Scale and rotation transformation invariants are typical features that are useful for identifying similarities between patent content images. Affine-SIFT (ASIFT) is the most recent improved version of these affine transformation algorithms and it is more reliable when compared to other algorithms.

Test Collections and Evaluation Measures

Patent retrieval systems are typically evaluated using both common information retrieval techniques and techniques specific to patent retrieval, such as:

(i) Recall (R) and Precision (P) at Top-K ranks.

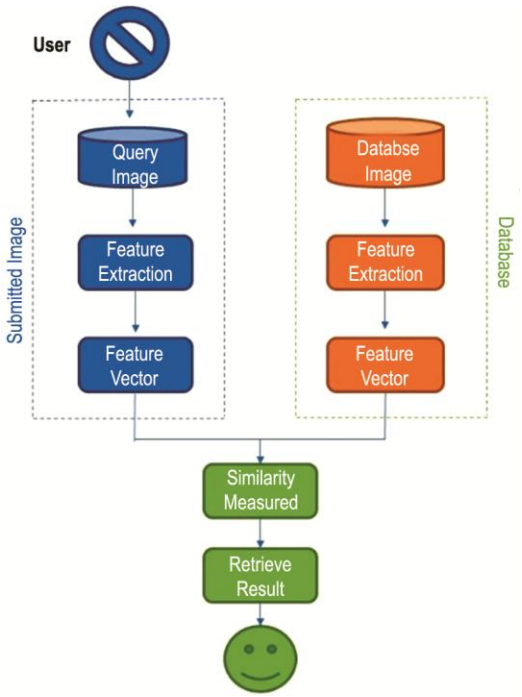

Fig. 4 — Content based image retrieval

(ii) Mean Average Precision (MAP), which places less emphasis on recall and generally favours early retrieval of pertinent documents.

(iii) The Normalized Discounted Cumulative Gain (nDCG), which promotes both the prompt retrieval of pertinent documents and the ranking quality of each document in turn.

(iv) For recall-focused tasks like PR, the Patent Retrieval Evaluation Score (PRES) was proposed. The rankings at which pertinent documents are returned provide an estimate of the user's review effort, which is what PRES focuses on as well as the overall system recall.

The relevant documents in other data-sets or database, collections like CLEF-IP 2012/2013, were compiled from citations that broke new ground discovered in search reports by examiners, as a result, these datasets are suitable for searches for invalidity and patentability, though invalidity searches require the use of non-patent literature too. The different research directions (Fig. 5) which were used for patent retrieval effectively are:

(i) Exploit patent structure for effective query formulation

(ii) Query Reduction / Query Expansion

(iii) Syntactic and / or Semantic properties of patent document

(iv) External knowledge base using semantic web.

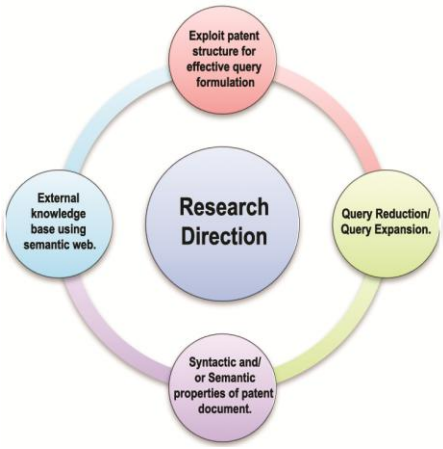

Fig. 5 — Research direction to inventor

Query Reformulation

The Query Reformulation techniques are the ones that are most frequently used for patent retrieval.[3] These techniques seek to the retrievability improvement of pertinent documents by converting the input query Q into Q through the contraction or expansion of Q terms.

Query Reduction (QR)

The term "Q terms" refers to a characteristic subset of terms taken from Q. This category includes terms that are taken directly from specific sentences or parts of the patent document, position-based methods are most frequently employed. any given greater matching weight compared to others. using a lexicon or stop-words list derived from IPC definitions for Q in IPC-based methods is another example of query reduction.

Query Expansion (QE)

Where characteristic terms that aren't in Q are taken out and combined with Q to create Q. The most common technique in this category is Pseudo-Relevance Feedback (PRF), where it is assumed that the top results of running Q are relevant and terms from these top results are used to expand Q terms. Other semantically based query expansion techniques operate by enlarging Q with terms that have comparable meanings, such as synonyms or hyponyms.

Hybrid

Here, Q is formed by subtracting unimportant terms and adding more pertinent terms to Q to create Q. The majority of techniques for query expansion can also be used for query reduction, where only terms that appear in the expansion list are retained and all other terms are pruned.

Keyword-based Methods

This group of methods retrieves pertinent documents by seeking exact correspondences between the search query(s) and the desired information. The closed vocabulary assumption, which underlies keyword search, states that Only terms found in the desired search data are used to create vocabulary.

Pseudo-Relevance Feedback (PRF)

It is generally accepted in QE with PRF that the top documents returned by user queries are that, Relevance feedback documents are relevant for learning expansion terms, which can strengthen search effectiveness. Our retrievability results, however, reveal a clear bias in favour of a particular subset of patents based on the queries retrieved from the query patents. These highly retrievable patents have the potential to skew the results, making a significant portion of patents either very poorly or completely unretrievable via any query.[3] A novel method for choosing relevant patents for PRF that identifies patents for PRF based less on the overall document similarity of the documents and more on their similarity to query patents more than a sub - set of terms was considered. This strategy's effectiveness is largely dependent on two elements. The most appropriate documents for PRF will be retrieved and enhance retrieval during QE, appropriate terms in the query patent must first be identified.

Semantic-Based Methods

Researchers have previously used semantic similarity to establish a connection between two key words.[5] To find a solution for a well-defined query in a text, Cui *et al.* used semantic expansion of the query. *Cui* used the query log to find a solution for the probabilistic correlation using the query and document terms. Large query logs aid in better retrieval outcomes. Wong *et al.* extracted phrases from the query using query expansion. These words and phrases are used to find similar words and phrases in the database. With the aid of various algorithms, the phrases are weighted. In order to save time and space, Subramanian *et al*, suggested using an improved stemming algorithm for data preprocessing, as well as links analysis techniques for information retrieval. The inbound and outbound links are evaluated for query expansion. Several similarity measures, including those based on stemming and language modelling, as well as purely lexical measures, are examined by Metzler *et al.* 2010.

Metadata-Based Methods

In addition to text, a great deal of non-text metadata and bibliographic data are also included in patents (e.g., tables, citations, formulas, classification, drawings, etc.). The advantages of using metadata features for CLIR include their language independence.[6]

Citation-Based

The most widely publicised method of metadata-based methods is the use of citation analysis to improve patent retrieval. The need for more thorough citation analysis was satisfied by the success of naively including citations from related patent applications as prior art.

Classification-Based

To enhance the efficiency of patent retrieval, these methods make use of classification information from both the patent topic as well as the retrieved documents.[7] Using IPC classification naively means keeping only documents with the same IPC classification code after filtering out all other documents that were retrieved as the topic patent at some level (for example, same subclass).

Interactive Methods

Patent retrieval that is interactive is unavoidable. Instead of improving retrieval performance, the current interactive methods for patent retrieval are more concerned with improving the organisation, use of structured, textual patent data and integration.[8] According to a recent analysis, performance would significantly improve if the user provided just one pertinent document manually is another motivation for the development of interactive methods for patent retrieval. Gains in performance from technology-assisted review. performance improvements through technology-supported review.[9] Investigate whether TAR protocols based on machine learning can be used for patent retrieval in fields like electronic discovery.

Conclusion

It is obvious that, in the absence of intentional domain customization and adaptation, it is impossible to directly apply in PR the efficient information retrieval methods used in fields like Web search. The literature review emphasises the unique nature of patents as well as how difficult it is to retrieve them. To increase the retrievability of patents, query expansion approach to false relevance feedback (PRF) were recognized. Based on an image query, Patent Image Retrieval (PIR) searches image databases for related patent images was approached. The goal of this paper is to suggest an improved feature extraction process for the patent image retrieval system.

References

1 Khode1 A & Jambhorkar S, A literature review on Patent information retrieval techniques, *Indian Journal of Science and Technology*, 10 (37) (2017) 4-7.

2 Bache R, Patent retrieval: A question of access, *World Patent Information*, 33 (2011) 345.

3 Shalaby W & Zadrozny W, Patent retrieval: A literature review, *Knowledge and Information Systems*, 1 (1) 2019.

4 Mogharrebi, *Retrieval System for Patent Images*, The 4th International Conference on Electrical Engineering and Informatics (ICEEI 2013), *Procedia Technology*, 11 (2013) 912 – 918.

5 Sharma P, Tripathib R & Tripathi R C, *Finding Similar Patents Through Semantic Query Expansion*, 11th International Multi-Conference on Information Processing-2015 (IMCIP-2015), *Procedia Computer Science*, 54 (2015) 390 – 395.

6 Ryley J F, Saffer J & Gibbs A, Advanced document retrieval techniques for patent research, *World Patent Information*, 30 (2008) 238–243.

7 Golestan Far M, Sanne S, Bouadjenek M R, Ferraro G & Hawking D, On term selection techniques for patent prior art search, *In: Proceedings of the 38th international ACM SIGIR conference on researchand development in information retrieval, ACM*, 1 (2015) 803–806.

8 (TAR) [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.*

9 Grossman M R & Cormack G V, Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review, *Richmond Journal of Law & Technology*, 17 (11–16) (2011) 153–162, 28.