Chapter 5
Evaluating Real Patent Retrieval Effectiveness

Anthony Trippe and Ian Ruthven

Abstract In this chapter we consider the nature of information retrieval evaluation for patent searching. We outline the challenges involved in conducting patent searches and the commercial risks inherent in patent searching. We highlight some of the main challenges of reconciling how we evaluate retrieval systems in the laboratory and the needs of patent searchers, concluding with suggestions for the development of more informative evaluation procedures for patent searching.

5.1 Introduction

Patent searching is a highly interactive and complex process often requiring multiple searches, diverse search strategies and careful search management [1]. There are different end-user requirements for different types of patent search, and simple performance-based measures of retrieval system functions are often inadequate in expressing the degree to which an information retrieval (IR) system might help conduct a successful search.

A particular characteristic of patent searching is the importance of the risk to which a company is exposed if a patent search is poorly conducted. Inadequate tools increase the likelihood of a poor search and increase the level of risk if a company proceeds on the basis of the search.

The claim from most IR evaluations is that measures of recall and precision, implicitly, calculate which system(s) are more likely to reduce this risk by performing more effective retrievals. Therefore, it is argued, we can be more confident about performing a good search with a system that has performed well in system trials. In this chapter we argue that this argument is naïve when considering real operational use.

A. Trippe (✉)
Patinformatics, LLC, 565 Metro Place S., Dublin, OH, 43017, USA
e-mail: tony@patinformatics.com

I. Ruthven
Department of Computer and Information Sciences, University of Strathclyde, Glasgow, G12 8DY, UK
e-mail: ir@cis.strath.ac.uk

M. Lupu et al. (eds.), *Current Challenges in Patent Information Retrieval*,
The Information Retrieval Series 37, DOI 10.1007/978-3-662-53817-3_5

Specifically we consider why recall and precision may give misleading interpretations on system performance, why we need to distinguish the characteristics of different types of patent search and where IR performance variability arises. A core theme in the chapter is the notion of risk: what risks are involved in patent searches, how these connect to measurements of recall and precision and how measurements of recall and precision may misinform rather than enlighten us as to system performance. We conclude with a discussion on how we might increase our confidence in IR system performance as measured in operational environments.

5.2 Types of Patent Search

Patent searches go by a variety of different names. Listing the most popular ones, you hear terms like: state of the art, prior art, patentability, validity, invalidity, clearance, freedom to operate, novelty and landscape (see Chap. 1). While there may be a large number of terms used to describe patent searches in essence, they boil down to four major categories upon which we shall concentrate in this chapter: state of the art, freedom to operate, patentability and validity.

Patent searchers traditionally use these types of descriptions to talk about the searches they perform for various clients whether they are from the legal department or the corporate strategy group. Before formally defining these types of search, it might be useful to think of these various types of searches in terms of the amount of risk they represent to the enterprise. Later we shall compare them to one another on precision and recall scales.

5.2.1 Patent Searches and Risk

We define risk as the amount of money that has already been invested in an innovation by an organisation pursuing a technological solution to a problem. As the amount of money invested by the enterprise increases, the importance of making good decisions about whether to continue funding the innovation and pushing it towards commercialisation also increases. With additional funding comes additional risk since the amount of money required to move from one step to the next in taking an innovation to market gets almost exponentially larger.

The pharmaceutical industry provides a perfect example of this concept of increased investment and risk. Early stage projects are expensive in terms of the time spent by the scientific teams in creating new drug entities and having those tested. These are sunk costs and are part of starting a pharmaceutical company in the first place. As new drug entities are discovered, however, decisions need to be made on whether they will be brought forward into what is first called a preclinical phase and then a succession of three human clinical trials. Each subsequent stage in this process becomes more expensive than the next as more people are involved in the trials, additional dosing schemes are employed and longer time periods are involved. As a company approaches a phase III clinical trial, the amount of money that will be invested is counted in the hundreds of millions of dollars and pale in comparison to the money that was spent generating a new drug entity and entering it into preclinical trials.

Since there is increased risk from substantially increased investment as a new drug entity moves from one stage to the next in the drug discovery process, companies have adopted a mantra referred to as ‘failing faster’. The idea being that if they can find mechanisms for discovering earlier in the process that a new drug entity is going to fail, then the company can save themselves a tremendous amount of money by learning as quickly as possible that this is the likely outcome. They cut down on later risk by identifying failure points earlier in the process before larger investments are made.

Analogies can be made to the world of patent searching from this example, and many companies follow a similar mantra that if they can discover potential legal impediments to future production earlier in the process, then they will save themselves money by changing course based on this knowledge. We can analyse the four major types of patent search by the risk involved.

State of the Art: This type of search is conducted in order to determine the prevailing technical knowledge in a particular subject area. A practitioner might be entering a new technical area and is interested in learning about the work that has already been done in this space. It is not uncommon for users to be interested in non-patent as well as patent documents in this case since the end goal is to have a thorough understanding of what the current knowledge is in a technical area of interest. People interested in technical or competitive intelligence will also be interested in these types of searches, and when they begin to analyse the details of the results they get, they will sometimes refer to these as landscaping studies. The sort of details a user can glean from these results are shifts in technology over time, interest in technology subcategories by company and who the subject matter experts in the field might be. State-of-the-art searches are typically done at the very beginning of projects before any investment has been made, and investigators are trying to determine if an innovation is worth pursuing for a number of reasons. The risk associated with these searches is low and this will have an impact, as we will see later on the corresponding need for precision and recall.

Patentability: This type of search is usually done in the legal context of determining if a new invention is eligible for patent protection and determining how broadly the claims for the new invention can be written. This type of search can cover both patent and non-patent literature and is typically looking for references that were published before the filing date of the invention in question. In the United States, inventors have up to a year from the first public disclosure of an invention to file a patent, so some searchers will go back an additional year with their searching to make sure they have found the best references. This is the type of search that will be done by an examiner to determine if they should allow a patent application to be granted.

Even though an examiner will do this search, it is important for the applicants to also conduct one since they will often have the time and resources to be more thorough than the examiner can be. It is also important since knowing the boundaries of the known references will help the attorneys drafting the claims to ask for the broadest coverage possible. Without knowing the scope of the known references, it is difficult for the attorney to know how broadly they can write the claims and still expect the examiner to grant a patent.

Patentability searches are done once an inventor has an idea and they have either reduced it to practice or they have a pretty good idea on how they are going to reduce the idea to practice during the preparation of the patent application. Investment has increased since the inventor has spent time discovering the idea and may have used additional time and money reducing it to practice. The total money spent, most of which is fixed costs, is still fairly low and thus the risk involved in this situation while higher than the stage when the state-of-the-art search was done is still low.

Freedom to Operate: Possibly the most specific type of patent search this particular one is country specific and only applies to in-force granted patents and their claims. A company will ask for a legal opinion on whether a product they are planning on shipping will infringe any existing patents before they launch. There is nothing offensive about this type of search since the interested party is not going to assert patents against anyone else; they are simply looking to make sure that they are not going to be infringing someone else's patents. A searcher in this case needs to identify the critical components of the product in question and search country-specific claims of in-force patents to see if any of them cover the product components in question. In most cases a great deal of money has gone into a product launch or can be involved with a successful product which is generating a great deal of revenue, so it is important for companies to know that they will be reasonably safe from future litigation before they make an even larger investment.

Some companies do freedom-to-operate searches reasonably early in the production cycle and follow up with them frequently to make sure the situation hasn't changed as they get closer and closer to market. These companies are following the 'fail faster' philosophy that was mentioned earlier since they recognise that it is better to know about potential legal issues before they make larger investments and involve higher risks. Other companies wait until the trucks are about to leave the warehouses and then conduct a freedom-to-operate search as a last item of their checklist before they go to market. At this point a great deal of time, money and effort has gone into the innovation and the amount of investment and risk is pretty high. On more than one occasion, companies have trucks filled with product that has been left in a warehouse because a last-minute freedom-to-operate search has come back with an in-force patent that could be used against the company later. Regardless of when these searches are applied, the risk is much higher than at the patentability stage and should be considered medium to high.

Validity: Validity search comprises the largest and most comprehensive of all patent searches. These searches are almost always associated with large sums of money and critical business decisions and as such need to be as comprehensive as possible. This search shares similar characteristics to the patentability search but is normally far more comprehensive since there is typically much more at stake when this sort of search is being initiated.

The object of the search is to identify prior art references which will allow a granted patent to be made invalid during a re-examination before the particular patent office of interest or during a court proceeding. Sometimes a company will also initiate validity challenges for patents that they are thinking of acquiring especially if they believe these patents will later be used in some type of litigation or another. On the flip side of this, a company who is provided with a cease and desist notice will often want to make the patents in question go away by finding invalidating prior art and then entering into re-examination. The prior art references in question can come from the patent or non-patent literature, must be available in the public domain and have to have been published prior to the priority filing date of the patent in question. In the United States there is a 1-year grace period on patent filings, so some searchers will look back an additional year when they search so they can be sure to avoid this type of situation.

Validity searches are conducted when an organisation has received a cease and desist order or is about to spend a significant sum of money on a purchase of one sort or another and due diligence needs to be performed in order to justify the transaction. Investment in this case either in the form of production costs and lost sales or in money to be spent on an acquisition is very high, and the corresponding risk to the groups making the investment is also extremely high. Since large sums of money are involved and the risk involved is so high, companies are willing to increase the resources made available to conduct these types of searches.

Summarising the searches on our risk continuum, we have state of the art followed by patentability, then freedom to operate and finally validity.

The amount of risk involved will have an impact on the resources that are made available to do the searching, and in turn this will have an impact on the precision and recall that will be expected in these searches. While risk is not the sole qualifier for precision and recall, there are cases where you have high risk but you do not need high recall per se; it is still useful to keep this in mind as we look at the requirements for these searches.

5.2.2 Risk and Recall

Looking at recall and thinking about a continuum, we come across an example where higher risk does not require higher recall. In the case of our highest risk search, validity, we also find that total recall is not necessarily required. In this type of search, it is only necessary to find *one* reference which predates the filing of the patent application in question that describes the invention. In practice most searchers will not stop when they find a single reference and will seek to be as comprehensive as possible, but strictly speaking it is not a requirement. Since there is a high risk, searchers will often seek higher recall to make sure there are contingencies in place and not rely on a single reference. These considerations put validity on the low-to-medium scale with regard to recall.

With patentability, the recall question will depend on who is doing the searching. In the case of an examiner, the recall will be the lowest of all the searches we are discussing since they will stop once they find a single reference which will enable them to disallow a claim. They can also take two references and combine them to disallow a claim, so they will stop if they find that combination. Patentability searches done by corporate searchers, however, are usually higher in recall since they are helping assist the attorney in deciding how broadly they can write their claims based on how much prior art is out there and how closely (we will do precision next) it matches the invention to be patented. Since the risk is still reasonably low, however, they will not attempt to achieve higher recall since they will reach a point of diminishing returns and making an investment to achieve it would not be economical.

State-of-the-art searches involve low risk, but you would like to achieve a reasonably high recall since the inventor is exploring an unknown area and they will spend time landscaping the area to increase their understanding. Economically speaking, recall is sacrificed due to the small investment being made at this point, and the bar for diminishing returns is pushed even lower since the expectation is that more comprehensive searching will be done once an actual invention has been discovered and when a product cycle starts.

For recall the top search is freedom to operate where a single missed patent can come back and be used for a cease and desist action. It is very important to find any and all patents that cover the elements of product to be brought forward to an attorney so they can make a determination as to whether the product will infringe on the patent in question. In order to conduct business, not just one patent can be found that an invention may infringe upon, but all of them need to be located in order to ensure that the company will not face future legal issues. These searches are referred to as freedom to operate for this exact reason.

So, looking again at our continuum and comparing recall, this time we have validity and patentability at the lower end of the scale, state of the art in the middle and freedom to operate at the high end. Recall does not correlate with risk necessarily in this comparison with the possible exception of freedom-to-operate searches.

5.2.3 Risk and Precision

Precision maps almost completely to our assessment of risk. State-of-the-art searches are sometimes called ‘quick and dirty’ since there is not much time invested in doing them and the results often have a large number of false positives contained in them. Also by its very nature, this search is exploratory and as such a high degree of precision is not required.

Patentability searches are typically more precise but by their nature are used to explore the boundaries of the prior art so that broader claims can be written to cover more aspects of an invention if warranted, so precision is important to cut down on the records that will need to be looked at but not essential. A number of false positives are expected and are part of the process.

Freedom-to-operate and validity searches both require a high degree of precision since very specific documents are required in each of these cases. With freedom to operate, the aspects of the produce must be covered in the claims of in-force patents from the countries of interest. The product must also use all elements of the claimed invention in order to infringe. Finding patents that meet this criterion is a tall order and requires high precision. Similarly, in a validity search, a precise search of the patent and non-patent literature is required to locate references which describe the exact invention covered in a later patent claim either by itself or in combination with another reference.

On the precision continuum, we have state of the art at the low end, followed by patentability and finally freedom to operate and validity.

Looking at each search by its characteristics, we can say state of the art is low risk with low precision and medium recall. Patentability is low risk with low recall and precision. Freedom to operate is high risk requiring both high recall and precision and validity having the highest risk and requiring high precision but able to get by with lower recall.

Looking at searches in this fashion, it is apparent that freedom-to-operate searches offer the most difficult challenge for IR researchers. The risks involved are also very high, so the expectations will be large and the reluctance to move away from established methods will be severe. Validity is also a difficult task since the risks are so high and the precision requirements so large. State of the art is where most systems work currently and do not necessarily provide much reward for the effort since they are low risk and are conducted with little in the way of investment. Patentability seems to be the sweet spot for IR research since it offers a reasonable challenge with a good opportunity for return since it is conducted during a stage where resources will be spent to address the issue.

Having outlined the challenges to the patent searcher in conducting a successful search, we now discuss some of the challenges IR researchers face in defining appropriate evaluation measures.

5.3 Limitations of IR Evaluation

As in other domains, the evaluation of the retrieval components of patent search systems focuses primarily on laboratory-style evaluation, and these evaluations are heavily shaped by the classical models of IR laboratory evaluation. As noted in Carterette and Voorhees (see Chap. 3 of the first edition article "Overview of

Information Retrieval Evaluation) early influential laboratory evaluations included studies such as the Cranfield I and II experiments, SMART evaluation and the in-depth evaluation and failure analysis of the Medlars search service [2] using small document collections. The experience gained from these studies has been incorporated into the creation of modern test collections where collection size has grown considerably since these early studies. The most widely used test collections come from the Text Retrieval Evaluation Conference (TREC) initiative [3], the Cross-Language Evaluation Forum (CLEF)[1] (which are discussed in separate chapters in this volume) and NTCIR.[2] The oft-stated values of test collection evaluations are the tightly controlled nature of the evaluation, the statistical rigour with which the evaluation test results can be analysed and the repeatable nature of the evaluation tests.

The value of IP systems in operational use, however, is influenced by more than the quality of the retrieval system itself and, as has been repeatedly demonstrated in operational tests in other domains, the contextual factors surrounding the *use* of a system (such as organisational concerns, training and experience of the searcher and time available to search) can strongly influence the end results of a search [4, 5]. This gap between real-life practice and laboratory rigour raises three important questions, which we shall examine in the remainder of this section.

1. Are laboratory evaluation measures misleading? Recall and precision are the standard measures for evaluating IR system performance. Although there are many ways in which we can use recall and precision to obtain evaluation measures, there are arguments for why they are poor measurements for end-user evaluations unless they are contextualised by other information. In Sect. 3.1 we examine some of these arguments and why they raise concerns for determining the confidence we can place in laboratory evaluation performance figures.
2. Are the results of laboratory evaluations sufficiently good at predicting real-life performance? That is, can the results obtained from a laboratory test of an IR system inform us of the potential value of a system in operational environments? In Sect. 3.2 we survey some recent work, which indicates a weak correlation between the performance evaluations of systems without user involvement and evaluations of systems operated by end users.
3. Are laboratory evaluations sufficient? Real-life evaluations incorporate factors that are usually eliminated from laboratory evaluations, such as the expertise of the searchers themselves. In Sect. 3.3 we examine some of these factors and outline their importance in reliably measuring system effectiveness.

[1] http://clef.iei.pi.cnr.it

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

5.3.1 The Potentially Misleading Effects of Recall and Precision

Patent search evaluation, similar to other retrieval problems, focuses primarily on recall and precision as measures of system effectiveness. These are long-held measures of retrieval quality and their tight hold on evaluation comes from their intuitive nature: how much of the useful information has my search retrieved (recall) and how much of the information that I have retrieved is useful (precision)? There is also a useful probabilistic interpretation of recall and precision: recall estimating the probability that a relevant document will be retrieved in response to a query and precision estimating the probability that a retrieved document will be relevant [6].

Most test collections are constructed using a generally accepted model referred to as the Cranfield model deriving from the Cranfield II tests [7]. A test collection that adheres to the Cranfield model will consist of a set of searchable objects, a set of information requests (or occasionally statements of information problems) and a list of which objects in the collection should be considered relevant for each information request. To ensure fair comparison between systems, a number of important assumptions are made. These include the assumptions that:

1. The topics are independent of each other.
2. All objects are assessed for relevance.
3. The judgements are representative of the target user population.
4. Each object is equally important in satisfying the user's information need.
5. The gathering of relevance assessment is independent of any evaluation that will use the assessments.
6. The relevance of one information object is independent of the relevance of any other object.

These assumptions are intended to ensure a fair and accurate comparison between estimates of system performance. The status of these assumptions has shifted over the decades of evaluation research since the original Cranfield model. Assumption 1 is generally adhered to in order to increase the diversity of the test. Assumption 3 is an attempt to ensure external validity of the experiment, i.e. that the results can be generalised to requests beyond those investigated within the test. The level to which this assumption matches most test collections is seriously under-investigated. Assumption 5 attempts to control the internal validity of the study: the assessments used to evaluate the system are not created by the people who designed the study, and therefore it is hoped that bias will not be introduced into the collection. Assumption 4 is a simplification of real search behaviour and many new test collections have graded relevance assessments to allow for more detailed measures of system effectiveness. However, the grades of relevance used often simply reflect amount of relevant material contained within objects rather than quality of relevant material. Assumption 6 is present in most test collections[3] although it is patently false—a system that retrieves duplicates or near-duplicate documents in favour of new and different relevant documents would not be seen as a better system by most users.

Assumption 2 is the assumption that has gathered most attention within the IR evaluation literature, particularly with the rise in test collection size. The early test collections contained small numbers of documents—the Cranfield collection contained only 1400 documents—and it was feasible for exhaustive relevance judgements to be made on the collection. For most collections this is not feasible: it has been estimated that it would take more than 9 months to judge an average size TREC collection for a single topic [7]. Not only is this expensive both in terms of time and resources, but over a protracted time period the criteria an assessor will use to judge a document for relevance could change, resulting in inconsistencies in the relevance assessments and therefore in the evaluation results. Indeed, Swanson [8] expressed this as one of his postulates of impotence—statements of what IR cannot achieve—namely, that it is never possible to verify if all relevant documents have been discovered for a request, as one can never examine all documents without unlimited resources while using a strict and static set of criteria for judging relevance. This is, of course, a real challenge for searches such as freedom-to-operate searches where the retrieval of all relevant documents is exactly what is required.

The reason that Assumption 2 has gathered so much attention is that exhaustive relevance assessment offers some guarantee that all relevant items have been identified, even if they do not linguistically match the user's query. That is, exhaustive assessments allow the identification of documents that conceptually match the query even if they do not match the user's choice of keywords.[4] Such assessments also allow for deep failure analyses of searches to ascertain why some search topics are more difficult for retrieval systems than others [9]. Such analyses are necessary, particularly with the current trend towards heavy averaging and aggregation of test results over large numbers of topics and collections. Several authors have argued against such approaches, particularly on the grounds that such tests are attempting to prove system hypotheses rather than disproving them. That is, experimenters are trying to prove a system works well rather than attempting to uncover when it will perform poorly. Such tests do not 'provide deep insights unless there is some degree of risk in the predictions' [10].

The current model for test collection—the pooling approach—is dependent on queries to create document assessment pools. Pooling compensates for exhaustive assessment by the inclusion of diverse systems and manual searching (see Sect. 2.2 of the first edition article "Overview of Information Retrieval Evaluation" of Carterette and Voorhees). The hope is that if we take sufficient care in sampling

[3] With the possible exception of INEX which does consider the relative relevance of sub-document units which may have overlapping content.

[4] Exhaustive query assessments also mean that we can assess the quality of the original query itself.

the documents to be assessed for relevance, we do not need to exhaustively assess the whole collection. The system-centred evaluation approach, therefore, argues that if we are sufficiently careful in selecting which documents are assessed and we evaluate on sufficiently large numbers of information requests, then we do not need to assess all documents in a collection.

The nature of test collection construction and the consequences of Assumption 2 are also important if we consider searching in operational environments. Test collection test results inform us of how well one system performs against another over a set of requests. Many studies have shown that the performance of any system across a set of requests is highly variable: systems will perform well for some requests and poorly for another. What IR tests cannot predict is how well a system will perform for a given request. This means, in operational environments, that the *searcher* must decide how well the system is performing for any given request. In many search situations, such variability might not matter; in patent searching it is more difficult to accept that some requests will be handled well and others not.

Blair and Maron [6] in one of most famous IR evaluation studies demonstrated that even experienced searchers can radically underestimate the proportion of relevant material obtained from an interactive search and that the quality of the searcher's queries can affect the *perception* of system performance. Although we can form intuitions about whether a system is returning relevant material, we cannot assess, simply based on the retrieved results, how much relevant material has been returned or how much remains to be retrieved. Blair and Maron in [6], and later [10], proposed four main reasons for the findings from their study:

1. Users often cannot predict which words are good at retrieving relevant material. In spite of detailed knowledge about the material with which they were involved, the searchers in their study could not identify useful search terms to retrieve important subsections of the database. However, they could consistently recognise useful information when it was presented to them. Common problems with querying included lack of knowledge of synonyms used in the unretrieved relevant material, poor handling of spelling mistakes relating to important terms and other oft-seen dilemmas in creating search requests.
2. The large size of the document collection meant that attempts to control precision—and hence make the result sets manageable—reduced the recall of searches. However, this results in the elimination of important relevant material from the search results.
3. Searchers can mistake document retrieval for data retrieval. That is, they describe the data they want to retrieve rather than the content of the documents they want to retrieve.
4. Overestimations of recall in laboratory tests give a false sense of security. In [10] Blair pointed out that poor laboratory tests can artificially inflate recall estimates. As noted above, test collection creators compensate for lack of exhaustive assessment by increasing the diversity of systems used to supply documents for assessments. The hope is that such diversity will lead to representative relevant documents being found. If the diversity is weak, then the recall figures can be artificially inflated because the relevant documents may be easier to find.

> Knowledge that one is using a good system can also give the searcher the perception that they are finding more of the relevant documents than they actually are.

What Blair showed was that, even by submitting variations of query terms adjusted through trial and error, as in a typical search session, the likelihood of a searcher finding a substantial proportion of relevant documents can be low, a finding that has been verified across a number of studies [10]. An explanation for this limitation is that the intellectual content of a document is difficult to represent automatically: a document can be about a topic without ever mentioning key terms or phrases that a user may expect to appear. In addition, the query terms chosen by the user may not discriminate between relevant and non-relevant documents, especially as the collection size grows [11]. A user searching for documents on a new subject may not select terms that are representative of the subject they are searching *and* that discriminate such documents from the non-relevant documents which share similar vocabulary. Consequently, not all potentially relevant documents will be retrieved through keyword matching techniques alone.

In a real search situation, a search can only estimate what is hidden (the unretrieved relevant documents) by what they have already found and by the quality of their attempts to find these documents. In [12] Blair argues that the latter is difficult to measure and searchers are often forced into intuitive reasoning about search strategies. One process known as 'anchoring' is of particular interest in searching. Anchoring is a psychological process in which people estimate unknown values (the quality of queries in our case) by starting from an initial value which 'may be suggested by a formulation of the problem'. If a particular query is seen as good, either because it retrieves relevant documents or the searcher believes it to consist of good indexing terms, then they will retain and modify the query, rather than attempt new queries, ones which may be better at retrieving different types of relevant material.

Blair and Maron's final point is also an important one for real search situations where the effort involved in conducting a search must be balanced against the cost of conducting a search: finding a number of relevant documents is not a sole indicator of good retrieval performance, as the proportion of relevant documents *missed* is not known unless it is quantified through other means. Swanson refers to this as the 'fallacy of abundance'—discovering a (substantial) number of documents about a request creates an illusion that little remains hidden [8]. Good precision, in particular, can give the false impression that the system has good recall.

There are two issues relevant for patent retrieval. Firstly, the degree to which recall and precision as measured in laboratory tests are actually informative of the likely performance in real situations. In the most challenging patent searches, simple measures of recall and precision may have little predictive power because what reduces company risk is not simply the ability to find relevant material but to have performed a comprehensive search. Very few system evaluations tackle the issue of how dependent system performance is on the initial request or how variable the system's performance is. Therefore, the end user's own expertise and judgement play a large role in the system's overall performance. Secondly, and as a consequence of the above discussion, we need to investigate the end user's abilities to make judgements about recall and precision in operational environments. Blair and Maron's studies indicated potential pitfalls about making such decisions in real-life settings, particularly when cost and time must be balanced against effort. As we will discuss in Sect. 4, there are ways in which we can estimate the skill of the person operating the system.

5.3.2 Predicting Performance from Laboratory Tests

One of the core claims for test collections, as noted, for example, in Sanderson and Zobel [13], is that the relative performance of systems from a test collection evaluation tells us something about how the systems will perform in operational settings. This is trivially true in extreme cases; a system that continually retrieves the wrong documents in a controlled test collection evaluation is unlikely to perform well in an operational setting. The test collection approach, typically but not always, concentrates on single retrieval runs. Some authors, such as Spärck Jones [14], have argued that this is not an issue; systems that perform well on one retrieval run will perform well in most retrieval situations and performance on single retrieval runs gives us an indication of how well a system will perform iteratively. However, single-run evaluation limits our ability to evaluate the effect of known aspects of how humans assess relevance, in particular dynamic effects such as the development of relevance criteria across a search [15] or the effect of the order of assessment [16].

However, the general claim that single-run retrievals are good estimates of overall system performance has not been convincingly demonstrated so far, partly due to the few comparisons in operational settings and partly due to the impact that user adaptation and interfaces have on the level of retrieval effectiveness of a complete system. What has been investigated is the degree to which laboratory tests and user tests align. This is **not** the same as tests in operational environments where many contextual factors will intervene.

Hersh et al. [17], who were one of the first authors to try direct comparisons between test collection and interactive experiments, show that results from a test collection do not necessarily follow to the interactive case because the interactive aspects of a system can interfere with the results. Their investigation also raises the question of what are *meaningful* differences between retrieval results: how much better does one system have to be over another in a test collection evaluation for us to be convinced that it is indeed a better system and are these differences the ones that are observable to users of the systems? Since Hersh and Turpin's paper, there have been a large number of attempts to shed light on the second question. The evidence is distinctly mixed. Kelly et al. [18], for example, showed that end users could distinguish or detect differences in retrieval performance but within tightly controlled environments where the users were forced to interact in specific ways. Hersh and Turpin's later results and Smith and Kantor's very robust study indicated, however, that users can compensate for the performance of poor systems [19] and, to a degree, undo the effect of good systems by raising their threshold for relevance [20].

Harter [21], for example, criticised the standard test collection model of evaluation because it ignored the variation in why relevance assessments are made for specific information requests. Relevance assessments in operational settings are heavily contextualised by the situation in which the assessments were made, and this context includes the person making the assessment.

Spärck Jones, in a later paper, also mentioned the importance of context and notes (of TREC in particular) ‘context is not embraced, but reluctantly and minimally acknowledged, like an awkward and difficult child. This applies even where explicit attempts have been made to include users (real or surrogate)’ [22]. Limited attempts to incorporate context within test collection environments have been attempted, notably in the TREC Hard and CiQA tracks, but these have typically related to the contextual information within the query rather than contextual factors which might affect the operational use of a system.

5.3.3 Are Laboratory Evaluations Sufficient?

Few evaluation measures and not those typically associated with test collections would take into account other factors that are important to users such as the validity of information, the ability of a searcher to understand the information retrieved, the source of the information or the searcher’s prior knowledge about a search topic [23]. Many studies (such as [24, 25]) have shown that, even for expert searchers, their confidence or prior knowledge in a search topic can affect their assessments of a document’s relevance: they will mark different documents as relevant, and different numbers of documents, independently of how those documents were retrieved. Voorhees, in a tightly controlled study, estimated the difference in opinion between assessors as around 35 %; Ruthven et al. [24, 25] indicated that differences also occur with individual assessors depending on their prior relationship to the search topic. Further, as noted above in Sect. 3.1, a searcher’s behaviour can strengthen the performance of a poor system or weaken the performance of a good system.

The question then arises as to what degree measures such as recall and precision obtained from laboratory studies actually help predict how good a patent search might be? If relevance assessments change depending on who is doing the assessment, then how much confidence can we have in evaluation measures based on relevance: if a different patent searcher conducted the same search, would we have different results? In operational environments, especially for searches with high risk, patent searchers can interact with each other to minimise the possible negative effects of individual variation in relevance judgements and search strategies.

However, as noted in Sect. 3.1, this places the emphasis for success onto the searcher and away from the system. A good set of evaluation measures would recognise and reward systems that offer support for end users in making challenging search decisions. The patent searches outlined in Sect. 2 are not simple searches; they are active processes where the end user must engage in a process of sense-making—understanding and interacting within information in complex ways to make a decision or recommendation. What makes a good IR system for this type of search behaviour is the ability of the system to make better sense of the search results and have more confidence in the accuracy of the outcome. This cannot be measured simply by performance evaluation but requires evaluating the process of searching. So how can we estimate the value of an IR system in helping successfully conduct a patent search?

In Sect. 4 we try to address this final question, building on the discussion in the previous sections, by outlining how we can gauge levels of trust in various parts of the IR process.

5.4 Evaluating Real Patent Retrieval Effectiveness

Any evaluation measure, implicitly or explicitly, carries a definition of success. This definition of what it means to succeed in an evaluation carries with it, in turn, the definition of what we see as the task of IR systems. In this chapter, we argue that the role of IR systems is to reduce overall risk; partly this is associated with measures of recall and precision (although simple measures may be too blunt), but the highly intellectual and interactive role of the patent search system (as a whole) needs to be incorporated into the evaluation.

One way of viewing IR evaluation is as a series of evaluation layers, each with distinct methodologies, metrics and questions. Lower evaluation levels comprise highly constrained, specific investigations on single system features; higher levels contain broader multifaceted investigations on the searcher *and* system. At the lower levels, for example, evaluations are typically on the algorithmic properties of system components and are run as performance tests conducted without human involvement. Higher levels will examine the interactive nature of the system to consider the degree to which the whole system supports an end user's information search. Appropriate metrics here will include both measures of the search products and the process of searching [26]. Product metrics, those that measure the end results of searching, may include aspects such as the number of relevant documents found, search satisfaction or time taken to complete a search. Process measures, on the other hand, consider how these products arose within a search and could include factors such as the ease of completing a search, the user understanding of the interface functionality, their increase in confidence in using the system and the use of system features.

As noted in Sect. 3, there are major differences between algorithmic evaluations and operational trials:

1. The effectiveness of a real patent search is dependent on the use of multiple systems and the searchers' ability to use them. Sections 3.2 and 3.3 outlined

some of the reasons why IR evaluations may not give us good predictions of how well a system performs in operational tests.

2. IR evaluation is based on generalisations. As noted in Sect. 3.1, IR evaluations tell us which systems are better for an average request. However, their performance across topics is very variable.
3. Individual estimates of recall and precision are affected by individual variation in how a searcher assesses relevance and what is returned by the system. It is far easier to reason about what is returned by a system than to reason about what is not returned.

Patent searching is a complex form of searching and one that involves multiple searches, collaboration with other people and heavy use of instinct and experience. So what types of evaluation are useful in understanding the success of an IR system for different types of patent searching? Arguably the success of any IR system is how well it supports the user in an information task, and measuring this will involve a number of different measures, some of which will be product based and some will be process based. However, as noted in Sect. 3, the ultimate purpose of IR tools within the IP process is to reduce risk by helping end users discover the required information or, alternatively, be reassured that certain information does not exist. Current laboratory evaluation measures do not help assess the degree to which an IR system has helped reduce this risk. Due to the variability in IR system performance, a user cannot guarantee any minimum level of performance for an *individual* search request. Nor can system designers assert, concretely, what level of confidence they should have in individual system components reducing risk because, as noted in Sect. 3, risk and recall/precision are not linearly related.

What we can try to develop are evaluation approaches that help estimate the confidence we should have in different system components. That is, how might we estimate what levels of trust we can have in parts of the retrieval process? If we have low levels of trust, then the end user needs to do additional work to compensate for lack of system performance.

5.4.1 Product-Based Measures for Evaluating Real Retrieval Effectiveness

Product measures are common in IR. Recall and precision can be used flexibly to give different estimates of system performance and different estimates are useful for different purposes. For a state-of-the-art search, reasonable recall is required and low precision perhaps tolerated, but debatably diversity of results is more important. Systems that artificially boost recall at the expense of missing important sections of the recall base could give the false impression that higher recall has been achieved. Systems may also be rewarded for retrieving some types of documents over others. In landscaping studies, it may be more useful for a searcher to have overview documents than narrowly focused documents. Calculating recall and precision over different document sets could be useful here.

For validity searches very precise results are required. Unlike state-of-the-art searches where we know there is material to be found but not sure what form it may take, in validity searches the question is whether the material is there to be found. In such a case, a useful evaluation metric may be final user confidence in the results of their search. A system that has a very high degree of topic variability (some queries are very successful, others very unsuccessful) offers little confidence in the performance on a new search. In such a situation, the searcher may have to expend more resources, time and cognitive, to complete the search but with little guidance from the system as to how effective the search has been.

Product-based metrics often focus on different systems with the same request; what they often fail to do is determine the variability of different requests on the same system. A useful product-based metric, particularly in light of the discussion in Sect. 3.1, is how variable a system performance is to the query formulation. High variation, particularly for best match systems, offers little confidence in the overall system performance and, again, increases the effort the searcher must expend on the search.

5.4.2 Process-Based Measure for Evaluating Real Retrieval Effectiveness

Process-based measures are useful for identifying the factors that lead to success and involve analysing the stages that lead to the end products of a search. In particular, for complex tasks where searchers may spend long periods of time on each search, process metrics are useful for identifying which search decisions are critical and which decisions need different types of system support.

Process measures are often difficult to develop and are subject to variation within the user population. However, process models can be used to (a) understand the processes of searching and (b) analyse success factors within each stage. An example of the latter is the University of Tampere's Query Performance Analyser [27], a tool for assessing how good a searcher is at the task of creating search requests. Such tools can help identify the relative contribution of the person conducting the search but also the contribution of the system to a successful outcome. Such knowledge could increase our confidence in the results of a search (in the case of high user and system abilities) or estimate what level of doubt we should retain after conducting a search. Understanding the process of searching within a professional domain like patent searching can also uncover the major sources of variation within patent searches and move towards correcting the sources of variation. Many disciplines use such process models to increase confidence in the overall process of completing tasks.

For high-risk tasks, such as freedom to operate, which requires both high recall and high precision, we could ask how individual searchers balance these requirements by the choice of search strategies and whether some strategies are more effective than others. Thus we can hope to move towards a more formal evaluation strategy for patent searching.

5.5 Conclusion

This chapter considers evaluating *real* retrieval effectiveness: retrieval effectiveness within an operational setting rather than in a controlled laboratory setting common to most IR evaluations. Deciding what to measure in evaluation is a crucial decision. It is worth reiterating the general point that any evaluation approach tends to distort what it tries to evaluate. Evaluation as an activity highlights some aspects of the phenomenon being studied and ignores others. As Hersh and Turpin [20] demonstrated, employing simple relevance metrics in user evaluations can give misleading results because simple metrics may ignore the factors that influence decisions. In this chapter, we have argued that retrieval system evaluation needs to provide a richer and more realistic account of the role of systems in reducing risk.

Each domain has its own challenges and presents new challenges to IR. IR researchers typically look at precision and recall simultaneously and measure their methods by how techniques stack up against both elements. When it comes to patent searching, it might be more productive to separate these functions so that they can be maximised independently. It has been demonstrated that risk, precision and recall do not follow the same linear path when discussing the various types of patent searches. Since this is the case, it might be more productive to begin with creating methods that produce high recall exclusive of precision. Once this is accomplished, the results can be ranked using different methods to improve precision and manage the way the results are shared with the searcher. It will likely be the case that different methods will be used to provide higher recall than those that can be employed to share records with higher precision. Instead of expecting a single method to do both, it would be useful to the patent-searching community if the process was done stepwise to maximise the value to the user.

It is received wisdom in the IR community that the variation between search requests is the greatest source of variation in retrieval system performance, and such variation is greater than the variation between end users. However, such claims are based on relatively artificial settings, and we still have relatively little empirical evidence on what components of a retrieval system are actually useful and the relative contributions of searcher and search system to overall success in patent searching.

We have, albeit briefly, suggested some evaluation directions that may help identify fruitful research directions in patent search evaluation. There are considerable challenges, particularly around issues of confidentiality, to be tackled, but if we are to move towards better evaluation procedures, then we need to be able to ask basic questions about the processes and decisions involved in operational patent environments.

References

1. Joho H, Azzopardi L, Vanderbauwhede W (2010) A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. 3rd Symposium on information interaction in context (IIiX '10)
2. Spärck Jones K, Willett P (eds) (1997) Readings in information retrieval. Morgan Kaufmann, San Francisco, CA
3. Voorhees EM, Harman D (eds) (2005) TREC: experiment and evaluation in information retrieval. MIT Press, Cambridge, MA
4. Ingwersen P, Järvelin K (2005) The turn: integration of information seeking and retrieval in context. Springer, Heidelberg
5. Hansen P, Järvelin K (2005) Collaborative information retrieval in an information-intensive domain. Inf Process Manage 41:1101–1119
6. Blair DC, Maron ME (1985) An evaluation of retrieval effectiveness for a full-text document-retrieval system. Commun ACM 28:289–299
7. Voorhees EM (2002) The philosophy of information retrieval evaluation. CLEF '01: Revised papers from the second workshop of the cross-language evaluation forum on evaluation of cross-language information retrieval systems, pp 355–370
8. Swanson DR (1989) Historical note: information retrieval and the future of an illusion. J Am Soc Inf Sci Tec 39:92–98
9. Voorhees EM (2005) The TREC robust retrieval track. ACM SIGIR Forum 39:11–20
10. Blair DC (1996) STAIRS redux: thoughts on the STAIRS evaluation, ten years after. J Am Soc Inf Sci Technol 47:4–22
11. Blair DC (2002) The challenge of commercial document retrieval, Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size. Inf Process Manage 38:273–291
12. Blair DC (1980) Searching biases in large interactive document retrieval systems. J Am Soc Inf Sci 31:271–277
13. Sanderson M, Zobel J (2005) Information retrieval system evaluation: effort, sensitivity, and reliability. 28th Annual international ACM SIGIR conference on research and development in information retrieval, pp 161–169
14. Spärck Jones K (2005) Epilogue: metareflections on TREC. In: Voorhees EM, Harman DK (eds) TREC: experiment and evaluation in information retrieval. MIT Press, Cambridge, MA, pp 421–448
15. Vakkari P (2000) Cognition and changes of search terms and tactics during task performance: a longitudinal study. RIAO 2004 (Recherche d'Information Assistée par Ordinateur), pp 894–907
16. Huang MH, Wang HY (2004) The influence of document presentation order and number of documents judged on users' judgements of relevance. J Am Soc Inf Sci Technol 55:970–979
17. Hersh WR, Turpin A, Price S, Chan B, Kraemer D, Sacherek L, Olson D (2000) Do batch and user evaluation give the same results? 23rd Annual international ACM SIGIR conference on research and development in information retrieval, pp 17–24
18. Kelly D, Fu X, Shah C (2010) Effects of position and number of relevant documents retrieved on users' evaluations of system performance. ACM Trans Inf Syst 28:1–9, 26, Article 9
19. Smith CL, Kantor PB (2008) User adaptation: good results from poor systems. 31st Annual international ACM SIGIR conference on research and development in information retrieval, pp 147–154

20. Hersh W, Turpin A (2001) Why batch and user evaluations do not give the same results. 24th Annual international ACM SIGIR conference on research and development in information retrieval, pp 225–231
21. Harter SP (1996) Variations in relevance assessments and the measurement of retrieval effectiveness. J Am Soc Inf Sci Technol 47:37–49
22. Spärck Jones K (2006) What's the value of TREC – is there a gap to jump or a chasm to bridge? ACM SIGIR Forum 40:10–20
23. Barry CL, Schamber L (1998) Users' criteria for relevance evaluation: a cross-situational comparison. Inf Process Manage 34:219–236
24. Ruthven I, Baillie M, Elsweiler D (2007) The relative effects of knowledge, interest and confidence in assessing relevance. J Doc 63:482–504
25. Ruthven I, Baillie M, Azzopardi L, Bierig R, Nicol E, Sweeney S, Yakici M (2008) Contextual factors affecting the utility of surrogates within exploratory search. Inf Process Manage 44:437–462
26. Borgman CL, Hirsh SG, Hiller J (1996) Rethinking online monitoring methods for information retrieval systems: from search product to search process. J Am Soc Inf Sci Technol 47:568–583
27. Sormunen E, Pennanen S (2004) The challenge of automated tutoring in web-based learning environments for IR instruction. Inf Res 9, paper 169

Springer

http://www.springer.com/978-3-662-53816-6

Current Challenges in Patent Information Retrieval
Lupu, M.; Mayer, K.; Kando, N.; Trippe, A.J. (Eds.)
2017, XIII, 455 p. 88 illus., 44 illus. in color., Hardcover
ISBN: 978-3-662-53816-6