World Patent Information 33 (2011) 345–351

Contents lists available at ScienceDirect

World Patent Information

journal homepage: www.elsevier.com/locate/worpatin

Patent retrieval – A question of access

Richard Bache

181 Copland Road, Glasgow, Scotland

Keywords:
Evaluation
Retrievability
Patent
Information retrieval
Recall-oriented domain
Corpus access
Non-Boolean
Hybrid systems

A B S T R A C T

The development of models and systems in Information Retrieval (IR) has been driven by the empirical measurement of effectiveness. However, in recall-oriented domains such as patent search where there is a significant cost of missing a relevant document, standard IR effectiveness measurement only reveals part of the truth. Since credible estimates of recall are not available, it is difficult to evaluate or design systems for this domain. Here, we propose a measure of corpus access, retrievability, and show using four large patent corpora that it can be used both to evaluate models for patent retrieval and also the corpora themselves for the ease with which a document can be retrieved.

1. Introduction

Information Retrieval (IR) has come a long way since the term was coined in the 1950s as a result of collaboration between librarians and the early computer scientists. The term 'Information Retrieval' is actually a misnomer since it is really about document location, whether that document is sitting on a bookshelf or, as more often today, within an electronic archive or on the World Wide Web. It is a technology that has broadened the originally intended user base of professional researchers and academics to anyone who can get online. Few users of Internet search engines have ever heard of the term 'Information Retrieval', yet they possess an implicit understanding of what it can achieve when they search for a suitable webpage without knowing its title, author or URL.

Over more than half a century, IR has developed models, a terminology and, perhaps most significantly, an empirical paradigm, which allows IR researchers and IR system designers to evaluate and compare competing systems. Naturally, IR has become an indispensable part of patent search. But we argue here that the empirical paradigm developed within the IR community cannot yet address the very nature of patent search and thus has difficulty in evaluating systems for its use. Put simply, most system (as opposed to user) evaluation is concerned with measuring how often users will encounter documents relevant to their needs.

Patent search is more concerned with ensuring that everything relevant has been found and often seeks to demonstrate that something (e.g. a document which invalidates a claim) does not exist. This is different from the prototypical IR task were a user seeks to find documents relevant to satisfying some information need. However, IR has provided the terminology and conceptual framework which allows us to describe this problem more clearly. Although IR evaluations have be geared toward measuring what is known as *effectiveness*, new research has attempted also to measure *access* which is a complementary rather than rival attribute of interest. Some early results show the potential of this approach are presented here as a means for evaluating and improving systems for patent retrieval.

The rest of the paper is structured as follows. Section 2 examines the Cranfield paradigm as now practiced by the Text REtrieval Conference (TREC) and discusses its limitations in the patent domain. In Section 3 we look at the essential structure of an IR system and review a selection of models that are used in IR generally. Section 4 explains the concept of access measurement and in particular how *retrievability* may be measured. In Section 5 we look at the problem of generating a very large number of queries which is necessary to perform this analysis. Section 6 presents some results of an evaluation of IR models with respect to measurement of access for 4 large patent corpora. In Section 7 we offer some conclusions.

Many of the results presented here also appear in [1], which was written for the IR research community. However, what we have attempted to do here is to present the material in a form accessible to patent searchers and other practitioners in the field. As a result, for example, the mathematical concepts are described verbally rather than by equations. Readers interested in a more formal approach are referred to [1].

2. Effectiveness and the Cranfield paradigm

Experiments in the 1960s on what we would now consider tiny collections comprising just hundreds of documents established the

E-mail address: richard.bache@gmail.com.

0172-2190/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wpi.2011.06.001

basic principles of effectiveness measurement used today. During the 1990s, TREC (Text REtrieval Conference) was set up as a mechanism for researchers to evaluate IR systems on considerably larger test collections. We now discuss the contribution this has made to IR and its limitations when used in the patent domain.

2.1. Precision and recall

An IR system is assumed to retrieve documents from some corpus. We assume that some user of an IR system has an information need, which he expresses as a query. This is submitted to the IR system under investigation and zero or more documents are retrieved – that is they are identified to the user as being potentially relevant. The actual task of fetching or downloading the document is considered outwith the scope of IR. We further assume that all documents in the corpus can be deemed either relevant or non-relevant to the information need, although what 'relevant' might mean is discussed below. There are two base measures (from which many other are derived) used to assess the *effectiveness* of the IR system with respect to the collection and query: precision and recall, which are defined as follows:

Precision: The proportion (or percentage) of retrieved documents that are relevant;

Recall: The proportion (or percentage) of relevant documents that are retrieved.

Many IR systems seek to rank documents by their computed relevance and so there is no clear distinct boundary between the retrieved and non-retrieved documents. The convention here is to set an arbitrary cut-off after, say, 10, 20 or 100 documents and measure, say, precision at 10, or recall at 100.

The issue of what relevance means in IR has been widely discussed [2–4]. Van Rijsbergen [5] gives an often quoted definition.

> *"A document is relevant to an information need if and only if it contains one sentence which is relevant to that need."*

However, in the patent domain relevance will have distinct meanings depending on the nature of the search task. Consider the following:

Novelty Search: A document is relevant if contains any information about prior art related to the invention.

Validity/Invalidity Search: A document is relevant if it contains any information that might invalidate one or more of the patent's claims.

Freedom to Operate: A document is relevant if it contains any claims which would restrict or prohibit the intended operations.

It is worth noting here that a patent searcher's view of an information need is somewhat different from most users of IR systems in that not only would he want to find relevant documents but, if none were found, would also wish to establish with some degree of confidence that no relevant documents actually existed. Indeed, often by finding one relevant 'kill' document this may be sufficient to halt the search.

2.2. Systematic measurement

The Cranfield 2 experiment [6] was the first attempt in the 1960s to evaluate competing IR systems in a scientific way. It aimed to create a situation where as many variables as possible were controlled. The experiment focussed on comparing indexing of documents but the methodology which it established could be generalized to other aspects of IR systems and more often now is used to evaluate scoring functions (both of these concepts are defined in the next section). To start with, a number of artificial but lifelike information needs were created and expressed as written descriptions. In TREC these are now referred to as *topics* although the term was not used at the time. From a set of topics, queries were devised. Then for all the documents in what was a small collection, each was judged for relevance against each topic. Such judgments require a human being that has some expertise in the area of the topic. Clearly any human judgment will be prone to subjectivity and error; however this was mitigated by making some assumptions about the nature of relevance. Specifically, it required assuming that whether a document is relevant or not is independent of whether any other documents in that collection are relevant.

Nowadays, from the set of topics and their associated queries, a corpus and the corresponding relevance judgments (now collectively known as a *test collection*), it is possible to evaluate any IR system in terms of precision and recall and other measures derived from them. To give a measure of the over-all effectiveness of an IR system, calculations are performed for many topics and an average taken. Many test collections are now available, such as those produced by TREC [7], and are considerably larger than the original Cranfield collections comprising typically hundreds of thousands of documents.

The fundamental difference between the modern large test collections and the original Cranfield collection is that exhaustive relevance judgments are not practical. For 100,000 documents and a set of 50 topics, the number of human judgments required is 5,000,000. Resources are simply not available to accomplish this task. As a result only partial judgments are provided, specifically for those documents which were highly ranked by one or more of the original IR systems under test. A vast majority of documents have no human judgment with respect to a given query. In any case, all that TREC evaluations seek to achieve is to compare the relative performance of systems rather than attempt to predict their actual precision when used in the field. As Zobel [8] points out, for the measurement of precision, there is no evidence that this underestimation unfairly favors one IR system over another. However, for recall, any estimate will lack credibility. Accurate calculation of recall requires the total number of relevant documents and the unretrieved relevant documents are simply not known. Imperfect measures of recall, based on only the known relevant documents, are used for the comparison of systems but for the recall-oriented domain, these are inadequate.

TREC organizes many different 'tracks', which relate to different domains of IR. Recently there have been the Chemical IR Tracks, CHEM '09 [9] and CHEM '10 [10]. Here the corpus comprises chemical patents and research papers. Instead of performing human relevance judgments, the track will consider papers and patents cited in the topic document (itself a patent) to be relevant. However, we note that this cannot amount to an exhaustive list of relevant documents. If it were, a validity/invalidity search would be trivial to accomplish.

The fact that recall is so difficult to measure means that development of models and systems has focused on the more measureable precision attribute. As we shall see, in many domains this is not perceived to be a problem, but in a small number of domains such as patent search, it has held back the development of suitable IR systems since there is no means to measure their efficacy with respect to these domains.

2.3. Precision and recall-oriented tasks

Tasks in many IR domains are precision-oriented in that achieving a high proportion of relevant documents within the top retrieved items is deemed more important than retrieving every relevant document. A good example is a web search by a casual user looking for, say, a recipe for chocolate cake. The user will want as many of the web pages returned to be actual recipes as possible. However, It would be reasonable to assume that the user would