World Patent Information 36 (2014) 4–15

Contents lists available at ScienceDirect

World Patent Information

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

Methods of ranking search results for searches based on multiple search concepts carried out in multiple databases

Alain Materne, Gershom Sleightholme*

European Patent Office, D-10958 Berlin, Germany

Keywords:
Multiple concepts
Search
Result ranking
Horváth–Materne ranking
Facet ranking
Pivot ranking

A B S T R A C T

If a patent prior art search produces several hundred results, it can be annoying when the best document turns out to be the last one. Ideally the most relevant search result should be brought to the top of the list. This article discusses the particular problems which searches for multiple concepts entail, explains what ranking is and compares some ways to rank, or reorder, search results. In particular, the article deals with a special technique which examiners at the European Patent Office (EPO) can use and which has been found to work well, especially when there are several search concepts and several technical fields to be searched. This ranking technique, called Horváth–Materne ranking or pivot ranking, is based on the assumption that the best documents will not only mention the search concepts in full-text databases, but will also mention at least some of the concepts in corresponding abstract databases. The more concepts present in the abstracts, the greater the probability that the document is relevant. In principle the technique could be used to rank results in any database where both full-text and abstracts are available. The technique can be broadened by including an automatic concept extraction.

1. Searching multiple concepts: introduction

In the past, searching for prior art in patent databases often involved just a single search concept in one CPC (Cooperative Patent Classification) class. For example, to search for a car seat belt which would release automatically if the car were accidentally driven into water involves a search in the CPC class B60R22/322 (automatic release of a seat belt in an emergency) with a single search concept 'water' and its synonyms. The class is very precise and the searcher or patent examiner can easily determine synonyms (e.g. water, damp????, humid+, moist????, river?, canal?, sea, lake?, harbo?r?, submer+, immers???) which would represent a complete search without generating much noise. The search concept could alternatively or additionally be represented by the CPC class B60R2021/0016 (type of accident: fall in water). In either case the search is straightforward.

These days, searching for prior art is often more complicated. Consider the example shown in Fig. 1 (from publication FR 2 954 611 A1, Faurecia Intérieur Industrie), which is a holder for a mobile telephone in a vehicle. When the mobile telephone 2 is placed against the front surface 8, a vacuum pump 25 removes the air from a pocket 17 so that the balls 20 inside it contact each other and form a rigid mass which grips the telephone. In this example four search concepts can be discerned: holder, vacuum, mobile telephone and vehicle. Some English synonyms for each of these concepts are shown in Table 1.

French and German synonyms are also searched. The synonyms need to be complete in order not to miss relevant documents, but as can be seen, they are not very precise and will certainly produce considerable noise. At the same time, in such a case as this the searcher can not be certain that all synonyms have been included. A further difficulty with searching multiple concepts is that there are frequently several relevant CPC classes, as shown in Table 2. Japanese, Chinese and Korean documents might not have a CPC class and need to be searched separately, e.g. using corresponding IPC or FI classes. In view of these problems, it has to be borne in mind when searching that the best prior art documents:

- might not mention all of the search concepts (e.g. use in a vehicle), or not with the synonyms used;
- might not be in the most obvious CPC class;
- might not have a CPC class at all (JP, CN, KR).

Typically, a thorough search will produce up to 1000 results, nearly all of which are irrelevant. To sift through the entire result set would take considerable time and be tedious. Instead, to see the best documents first, the 1000 results need to reordered, or ranked.

* Corresponding author.
E-mail addresses: amaterne@epo.org (A. Materne), gsleightholme@epo.org (G. Sleightholme).

0172-2190/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wpi.2013.09.003

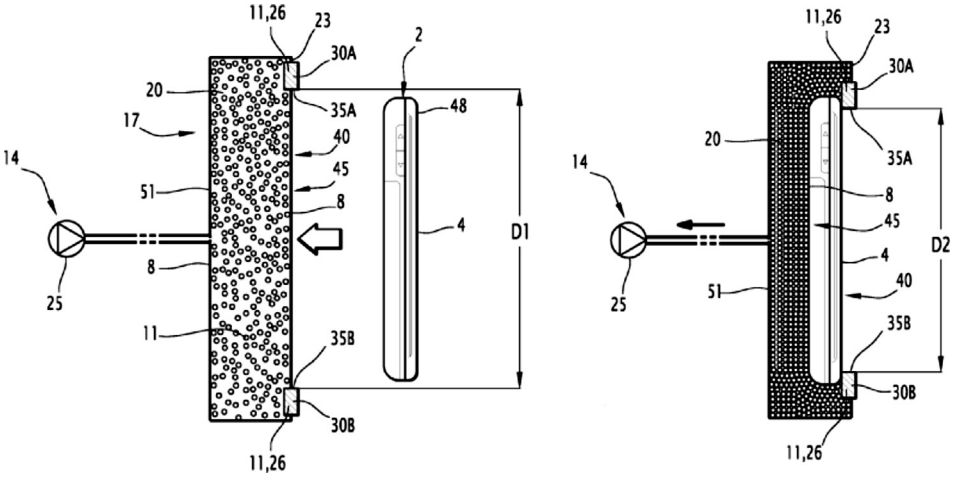

Fig. 1. Multiple search concepts (vacuum holder for a mobile telephone in a vehicle).

2. What is ranking?

Ranking is reorganising search results so that the most relevant information appears higher in the list. It can be carried out in a number of ways—based on different methods and criteria. In other words, put very simply, ranking can be defined as:

- reordering (i.e. the number of documents does not change as a result of ranking, the documents are just rearranged)
- of a result set (i.e. a search has already been carried out and, if applicable, search concepts combined using 'and' or 'or')
- based on certain criteria

It turns out that the criteria used for ranking are critical: how the ranking is carried out makes an enormous difference in whether the best documents really come out on top or not. Some currently used ranking criteria are outlined in the following sections.

2.1. Frequency ranking (occurrence ranking)

One of the simplest criteria for ranking is the number (frequency) of occurrences of the search terms. For instance, in the example with the releasable seat belt, documents in the result set are reordered according to how often 'water' or any of its synonyms occur in the document. This type of ranking is implemented at the EPO by a command '..rank', ('..' characterizes any instruction or script started on the EPOQUE search engine) in which a basic highlight counting of searched keywords is performed. For single search concepts, this often produces acceptable results as E. Nijhof observed [1]. Any use with more than one concept without caution will lead to unpredictable results, as can be seen from the following example.

Referring to the hypothetical results shown in Table 3, document D2 has the highest score (rank 1), but only mentions concepts 3 and 4 once, possibly in a discussion of the prior art. According to this ranking criterion, D2 would therefore be displayed first, although it is not relevant. Refinements can be made to compensate for document length or to apply a weighting to certain concepts, but still the results are not always as desired. Furthermore, the computer processing time can be considerable.

2.2. Horváth ranking (facet ranking)

A technique which has been used successfully to rank multiple concepts for over a decade at the EPO is based on the assumption that the more concepts, also defined as facets by S. Ranganathan [2], which are present in a document, the greater is the probability that the document is relevant. The technique is also known as Horváth ranking after the EPO examiner who developed it. Result sets are formed by making 'or' combinations of the search concepts; the documents are then ranked by sorting them into different subsets, the first subset containing documents having all search concepts, the next subset containing documents with one fewer concept, and so on, as shown in Fig. 2. At the EPO a script can be used which generates the sets of Boolean combinations, see also [3].

The concepts can be word queries in full text, abstracts or titles, or classes, or combinations as this type of ranking does not distinguish between data types. Good documents may still appear in one of the lower subsets even if one of the concepts had been expressed differently - facet ranking is tolerant with documents (Horváth)—thus resulting in a high level of precision according to H. Iyer [3]. The different subsets in facet ranking also correspond in a way with novelty and inventive step objections: the missing concept might be obvious.

When some particular concepts need to be emphasized in the results of ranking, they can be forced into the output by setting these concepts as "musts". A forced concept is no longer defined as a facet but is converted to the status of a must.

The number of musts can be selected so as provide an additional facility on top of the facet concepts. Their maximum number is set as a balance to the number of usable facets.

Table 1
English synonyms for four search concepts.

1	Holder, attachment, bracket, ...
2	Vacuum, suction, low pressure, ...
3	Mobile/portable phone, electronic device, smart phone, PDA, ...
4	Vehicle, car, automobile, transport ...

Table 2
Possibly relevant CPC classes.

B60R11/0241	Holders for telephones in vehicles
B60R11/0235	Holders for flat screens in vehicles
H04M1/04	Supports for telephones
F16M13/00	Supports in general
G06F1/16(?)	Constructional details of data processing equipment
A45F5 (?)	Holders for hand articles