Relevance Ranking in Patent Databases. Is It Relevant?†

Edlyn S. Simmons

Hoechst Marion Roussel, Inc., 2110 E. Galbraith Road, Cincinnati, Ohio 45215-6300

Received August 12, 1998

Patent databases grew up with the traditional online search services, where Boolean logic and proximity operators are used to produce a set of answers in reverse chronological order, some of them relevant, some false drop. Newer search engines are based upon more complex algorithms that automatically combine traditional operators with occurrence counts to generate sets of answers ranked according to their relevance. The underlying assumptions of relevance ranking may not be valid for patents, however. Because patents use generic language in preference to specific terms and rely upon drawings and chemical structure diagrams in preference to verbal descriptions, algorithms that rank answers on the basis of word frequency and proximity may not identify the most relevant patents. The wide availability of U.S. patent databases searchable with various search engines facilitates comparison of the capabilities of relevance ranking algorithms.

Patent searching is now in its fourth technological era. The earliest patent search system was manual. Hierarchical classification systems were devised by patent offices, and one or more classification codes were assigned to each patent. Paper copies of the patents were arranged in stacks according to their classification code, and a searcher picked up a stack of patents covering the subject of interest, looked at each one, and when a patent was recognized as relevant, picked up a pencil, and made notes on a piece of paper. Patents could also be searched with a printed list of index terms or a simple card file produced by patent offices, commercial indexing organizations, or private firms. A major breakthrough in patent searching in the early 1950s was the IFI dual dictionary, which contained two copies of the index that could be searched side by side.[1] Only patents indexed under both terms would be sufficiently relevant for the searcher to take the time to view the actual document.

The second era was mechanical. With the availability of the early computer in the 1960s, it became possible to index larger numbers of concepts and to break chemical structures down into substructural units.[2,3] Information about the content of each patent was described by codes corresponding to positions on a punched card, and the card was perforated in each of the appropriate positions. The entire collection of cards was passed through a mechanical card sorter with needles or spikes corresponding to the desired codes, and cards with holes corresponding to all of the needles dropped into a collection area. This system could be repeated so that the Boolean "or" and "not" were accommodated as well as the Boolean "and". The searcher reviewed patent documents identified by the cards or abstracts printed on the card itself, and when a relevant patent was discovered, made notes on paper. Because the number of concepts that could be encoded was limited by the number of spaces on the card, many of the patents retrieved by this system were not, in fact, relevant and were labeled as "false drops".

The third era began with the emergence of online search services in the 1970s. Time sharing on remote computers freed individual companies from the necessity to maintain databases in-house. Because the amount of computer capacity available was limited, data was encoded in the same ways as in the era of mechanical card sorters, but the data was stored on tapes or disks and results of Boolean search strategies could be manipulated more readily by the computer and printed out at the searcher's site. The increasing size and capacity of the computers permitted the text of the documents to be added to the database and searched alone or with the coded information. Proximity operators made it possible to analyze text more precisely than Boolean operators. During the 1980s topological indexing of chemical structures made it possible to search for chemical substances with far greater precision than was possible with fragmentation codes.[4] As always, the documents or their surrogates were reviewed by a human, and relevant documents were separated from false drop.

In the new era of the 1990s, larger and more powerful computers can store the full text of millions of documents, and the text can be searched rapidly with the aid of sophisticated relevance ranking algorithms. In theory, at least, the most relevant documents are printed at the top of the list of retrieved documents, and the false drop falls to the bottom of the list. The searcher should not have to look at the lower ranked documents, because the computer will have reviewed the content of the documents.

But what exactly is relevance? The American Heritage Dictionary[5] defines relevance as "The capability of an information retrieval system to select and retrieve data appropriate to a user's needs" and ranking as "To give a particular order or position to; classify". Relevance ranking, then, should be the capability of an information retrieval system to classify data in order of its appropriateness to a user's needs. This is simple enough if it is done by a human being who knows which documents are more relevant and which are less relevant after reading them. But how is it possible for a computer algorithm to judge the relative relevance of documents before we read them?

† Presented at a symposium on "Patent Searching as the Millennium Nears: New Tools, New Techniques", for the Division of Chemical Information at the American Chemical Society National Meeting, Dallas, TX, March 30, 1998.

10.1021/ci9801421 CCC: $18.00
Published on Web 04/17/1999

- QPAT-US, http://www.qpat.com
  - Full text and front page/abstract databases
  - Personal Library Systems search engine
- USPTO, http://www.uspto.gov/patft/index.html
  - Front page/abstract
  - CNIDR search engine
- IBM Patent Server, http://www.patents.ibm.com
  - Front page/abstract and claims
  - Verity search engine

Figure 1. Relevance ranking search systems for U.S. patents.

Relevance ranking software was developed on the basis of certain assumptions about the structure of documents and the needs of searchers. Search engines that rank documents according to relevance are based upon the science of linguistics, the study of the nature, and structure of human speech. Their algorithms sort documents according to a number of different considerations:

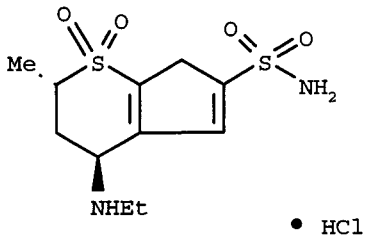
- The frequency with which a particular word occurs in spoken or written usage, which obviously differs widely from language to language, and the frequency with which a particular word occurs in the database being searched. Words that are used frequently are given a low weight, while those that are used infrequently are given a high weight.
- The frequency with which a particular word occurs in a single document. A term that is repeated often is likely to be central to the meaning of the document.
- The number of words in the document. The number of occurrences of search terms in a document is weighted relative to the length of the document.
- The placement of the word in a document. A word that appears in the title may be considered to be more important than a word that appears somewhere in the middle of the text.
- When more than one word is used as a search term or when a single search term occurs in a document more than once, the proximity of the search terms may be considered.
- When more than one term is used in a search, the number of different search terms found in a single document may be considered.

Each search engine has its own relevance ranking algorithm. Some search engines allow the searcher to specify Boolean operators and proximity operations, which may override the relevance ranking algorithm or which may supplement it, changing the ranking of documents retrieved by a search.

We usually associate relevance ranking with the search engines that find universal resource locators on the worldwide Web, but there are also search engines that search databases rather than websites.[6,7] There are several United States patent databases on the worldwide web that use relevance ranking algorithms and allow the searcher to use multiple search terms with Boolean logic.[8−11] Three of them, Questel-Orbit's QPAT-US, the IBM Patent Server, and the U.S Patent and Trademark Office's U.S. patent site, were used or the present comparison (Figure 1) QPAT covers U.S. patents from 1974, when the U.S. Patent and Trademark Office began storing all patents in electronic form, the IBM site has patents from 1971, when the USPTO began storing incomplete collections of patents electronically, and the USPTO's site includes patents from 1976 to the present. At the time this was written, QPAT-US had both a full text database and a database limited to front page information and abstracts; the IBM site had the first page data and claims, and USPTO site had searchable bibliographic and abstract data only. There are also relevance ranking commands on some of the traditional search services. DIALOG's TARGET command was used to rank search results from the PATFULL databases, files 652, 653, and 654, and STN's FOCUS command was used to rank output from the USPATFULL file, both full text databases.

DIALOG's TARGET command can be performed directly on a search statement or subsequently on an answer set, in command or menu mode. It returns only the top 50 ranked documents and ranks them according to a percentage relevance. STN's FOCUS is performed on an answer set, ranking all documents in the set and listing them in order of relevance. QPAT runs on a Personal Library Systems search engine; it returns all search results in relevance ranked order and assigns a numerical rating to each document. The USPTO database runs on a CNIDR search engine; it provides relevance ranking as an option and assigns a rating between 1 and 100. IBM Patent Server runs on a Verity search engine; it returns only the top ranked documents, with a default of 50 and a maximum of 200, and provides a percentage relevance ranking.

A good argument can be made that when the needs of the users can be accurately defined by the terms used for searching, relevance ranking algorithms can promote the most significant documents to the beginning of the list. But even if the computer can actually sort the documents it retrieves in order of their relevance to the hypothetical user, is it possible that the rankings assigned to patents are as accurate as those assigned to less specialized documents? Patents, after all, are not constructed in the same way as newspaper articles and stock analysts' reports. They have specialized text fields, such as claims and cited references. Much of the language is generic rather than specific. Much of the information in patents is expressed graphically rather than linguistically. Patents in all technologies have formal drawing pages. Chemical patents have chemical structures and reaction schemes. Biotechnology patents have protein and nucleic acid sequences. And many patents describe multitudes of chemical compounds by means of Markush structures. Special characters such as Greek letters and mathematical symbols are sprinkled liberally throughout patents, and subscripted and superscripted letters are used routinely to designate Markush groups. All of this is added to the fact that patents cover the entire range of technologies and include jargon and specialized terminology from all of them. Furthermore, the needs of the user of a patent database are often different from the needs of a user of business information or news. In addition to searches for general technical information, patent searchers do patentability and validity searches, where generic and suggestive disclosures are as significant as specific teachings, and freedom to operate searches, where the scope of the patent claims is of



(4S-Trans)- 4H-4-(Ethylamino)-5,6-dihydro-6-methyl-thieno[2,3-b]-thiopyran-2-sulfonamide 7,7-dioxide monohydrochloride,
OTHER NAMES:
Dorzolamide hydrochloride
L 671152
MK 507
Trusopt

Figure 2. Dorzolamide HCl.

```
(CLAIMS/RRX)
PATENT              US4797413   89.01.10
PATENT ASSIGNEE     MERCK & CO INC
ACTION CODE         92.11.24  REEXAMINED  CERTIFICATE B14797413,SEQUENCE 1852nd
                    REQUEST - 90/002682, Merck & Co, Rahway NJ, US(92.03.23)
                    CLAIM - AS A RESULT OF REEXAMINATION, IT HAS BEEN DETERMINED THAT:
                    The patentability of claims 1-13 is confirmed.
ACTION CODE         97.05.20  EXTENDED    1,233 Days (97.04.21)
```

Figure 3. Dorzolamide is claimed in U.S. 4,797,413.

Set #	Expression	Results
S1	dorzolamide	5

Score	Patent	
208	US5574176	Synthesis of an intermediate for the preparation of 5,6-dihyd...
206	US5688968	Enantioselective synthesis of 5,6-dihydro-(S)-4-(ethylamino)-...
202	US5441722	Short synthesis of 5,6-dihydro-(S)-4-(ethylamino)-(S)-6-[C.su...
200	US5605906	Cannabinoid receptor agonists
200	US5532237	Indole derivatives with affinity for the cannabinoid receptor

Figure 4. QPAT-US full text search "Dorzolamide".

paramount importance and other parts of the patent specification are irrelevant.

To find out whether relevance ranking might meet the needs of patent searchers, some exemplary searches were made using two queries. One of the topics is the drug called dorzolamide hydrochloride (Figure 2). Dorzolamide, a carbonic anhydrase inhibitor, is the active ingredient in Merck's Trusopt eyedrops, which are used for treating glaucoma. The other topic is the use of zinc compounds for treating the common cold, which is an infection caused by a rhinovirus. The keywords used for that search were (Zinc or Zn), cold, (rhinovirus or rhinoviral), and (antirhinovirus or antirhinoviral), appropriately truncated. Relevant patents were recognized by reading the titles and other text in the retrieved patents. No assumptions were made about the kind of information the hypothetical searcher was looking for, e.g., a comprehensive search for all references or a freedom-to-operate search for a specific pharmaceutical formulation, and the patents that were retrieved were not read, so no stringent standard for relevance can be applied to the results.

One basic real-world rule cannot be ignored in assessing relevance, however. The patent that claims a new compound, per se, is always extremely relevant to a search covering that compound. Therefore, in searches for dorzolamide that are not limited by other concepts, this patent should be highly ranked. Since dorzolamide is approved for sale in the United States and its patent has been extended, the patent that claims dorzolamide is easy to identify. The patent is U.S. 4,797,-413, which was originally granted in 1989. The patent was also reexamined. Figure 3 illustrates part of the record from the Claims Reassignments and Reexaminations database. By any objective standard, there is no more relevant dorzolamide United States patent.

The first test of relevance ranking was a simple search of QPAT for the generic name of the drug dorzolamide, summarized in Figure 4. There were five postings for the

QPAT-US full text search. "Dorzolamide or dorzolmide"

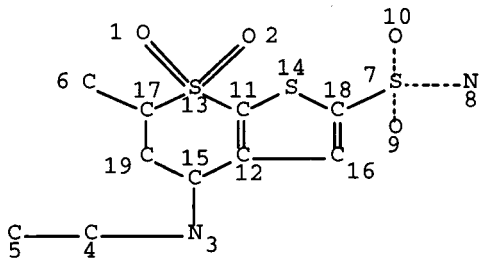
Set #	Expression	Results
S2	dorzolamide DORZOLMIDE	5
S1	dorzolamide	5

Score	Patent	
402 (202)	US5441722	Short synthesis of 5,6-dihydro-(S)-4-(ethylamino)-(S)-6-[C.su...
208	US5574176	Synthesis of an intermediate for the preparation of 5,6-dihyd...
206	US5688968	Enantioselective synthesis of 5,6-dihydro-(S)-4-(ethylamino)-...
200	US5605906	Cannabinoid receptor agonists
200	US5532237	Indole derivatives with affinity for the cannabinoid receptor

Figure 5.

```
FILE 'REGISTRY' ENTERED AT 14:26:08 ON 23 MAR 1998
COPYRIGHT (C) 1998 American Chemical Society (ACS)

=> structure
ENTER NAME OF STRUCTURE TO BE RECALLED (NONE):120279-96-1
ENTER (DIS), GRA, NOD, BON OR ?:.
```



```
L1    STRUCTURE CREATED

L2             24 SEA FAM FUL L1

FILE 'USPATFULL' ENTERED AT 14:29:24 ON 23 MAR 1998
CA INDEXING COPYRIGHT (C) 1998 AMERICAN CHEMICAL SOCIETY (ACS)

=> s l2
L4             7 L2

FILE 'MARPAT' ENTERED AT 14:27:44 ON 23 MAR 1998
COPYRIGHT (C) 1998 American Chemical Society (ACS)

=> search l1 css ful
L3             6 SEA CSS FUL L1
```

Figure 6. Chemical structure search. STN Files Registry, USPATFULL, and MARPAT.

word "dorzolamide". The patent numbers and their ratings are shown in the table, with the titles in the truncated form provided in the database results display. The ratings cluster just over 200. QPAT has an "EXPAND TERMS" button allowing the searcher to review the index for related search terms. One of the options, the alphabetical index of terms appearing in QPAT, showed that there was a misspelling of "dorzolamide" somewhere in the database. When the misspelled term "dorzolmide" was added it to the search using the default Boolean operator, which is OR, the same five patents were retrieved. As shown in Figure 5, however, the rankings in the results list were changed. The rating for U.S. 5,441,722, had risen from 202 to 402, and it moved from third position to first. The search engine had rated this patent as more relevant than the other patents because it had two search terms rather than one, and because the second term, the typographical error, had only one posting in the database. The algorithm is obviously mistaken about the relevance of rare terms; rare terms are often misspellings, and typographical errors are not indicators of relevance.

One of the best ways to do a search for a specific compound like dorzolamide is to search in a database that indexes chemical structures from patents. A family search was done in the CAS Registry file on STN, as shown in Figure 6, and retrieved dorzolamide and a cluster of salts, isotopes, and enantiomers. The Chemical Abstracts Service

"Dorzolamide or L 671152 or MK 507 or Trusopt"

Patent No.	Title	Reg.No.	Marpat	Focus Rank	Target Rating	QPAT Rating
US5620970	Topical ophthalmic carbonic anhydrase inhibitor formulations	X				
US5091409	4-Alkylamino-6-(C.sub.3-5–hydro-carbyl)thieno[2,3-B]thiopyran-.	X				
US4968815	Synthesis of (S)-3-(thien-2-ylthio)butyric acid analogs	X				
US4968814	(S)-Alkyl 3-(thien-2-ylthio) butyrate and analogs and synthesis	X				
US4797413	**Thieno thiopyran sulfonamide derivatives, pharmaceutical ...**	**X**				
US5441722	**Short synthesis of 5,6-dihydro-(S)-4-(ethylamino)-(S)-6-[C.sup.3H.sub.3]**	**X**	**X**	**1**	**99%**	**202**
US5565345	Process for preparing thienothiopyran derivative			2	46%	320
US5688968	**Enantioselective synthesis of 5,6-dihydro-(S)-4-(ethylamino)-(S)-6-**	**X**	**X**	**3**	**42%**	**206**
US5723628	Process for preparing carboxylic acid derivative			4	43%	336
US5574176	Synthesis of an intermediate for the preparation of 5,6-dihydro-(s)-			5	43%	208
US5494901	Topical compositions for the eye comprising a .beta.-cyclodextrin ...			6	42%	336
US5418225	Topical compositions for the eye comprising a .beta.- cyclodextrin...			7	42%	336
US5545651	Imidazole 5-position substituted angiotensin II			8	40%	304
US5512681	Angiotensin II receptor blocking imidazolinone derivatives			9	40%	304
US5424450	Angiotensin II receptor blocking imidazolinone derivatives			10	44%	312
US5411980	Substituted triazolinones, triazolinethiones, and triazolin- ...			11	40%	312
US5395844	Imidazole 5-position substituted angiotensin II			12	40%	312
US5389635	4- O or 5-Heterocyclic substituted imadazoles as as angiotensin-II- ...			13	44%	312
US5376666	Angiotension-II receptor blocking, azacycloalkyl or azacycloalkenyl ...			14	40%	320
US5310929	Prodrugs of imidazole carboxylic acids as angiotensin II receptor ...			15	40%	320
US5308846	Quinazolinones			16	40%	320
US5260325	Angiotensin II receptor blocking tertiary amides			17	40%	312
US5256667	Quinazolinones and pyridopyrimidinones			18	40%	312
US5219856	Angiotensin-II receptor blocking, heterocycle substituted imidazoles			19	40%	312
US5202322	Quinazolinone and pyridopyrimidine a-II antagonists			20	40%	312
US5175164	Angiotensin II antagonists incorporating a substituted indole			21	40%	320
US5605906	Cannabinoid receptor agonists			22	40%	200
US5554625	Substituted biphenylmethylimidazopyridines			23	40%	200
US5532237	Indole derivatives with affinity for the cannabinoid receptor			24	40%	200
US5266583	Angitotensin II antagonist			25	40%	200
US5264447	Angiotensin II antagonist			26	40%	200

Figure 7. Dorzolamide patents retrieval through CAS registry numbers, Marpat, and full text files as ranked by STN's Focus, DIALOG's Target, and QPAT-US.

S3: 12219 (COLD OR RHINOVIR? OR ANTIRHINOVIR?) AND (ZINC OR ZN)

Target Rank	Target Rating	Patent No.	Title	QPAT Rank
1.	99%	5,629,099	Alloying-treated iron-zinc alloy dip-plated steel sheet excellent inpress-formability and method for	>50
2.	87%	5,409,905	Cure for commond *cold	4
3.	84%	5,582,817	Remedy for dermatopathy and metallothionein inducer	>50
4.	83%	5,316,652	Method for manufacturing iron-*zinc alloy plated steel sheet having two plating layers and excellent	>50
5.	82%	5,464,596	Method for treating waste streams containing *zinc	>50
6.	81%	5,002,837	*Zn -Mg alloy vapor deposition plated metals of high corrosion resistance,as well as method of	>50
7.	78%	5,135,817	*Zn -Mg alloy vapor deposition plated metals of high corrosion resistance, as well as method of	>50
8.	77%	5,453,111	Method for separation of metals from waste stream	>50
9.	73%	4,897,317	Corrosion resistant *Zn -Cr plated steel strip	>50
10.	71%	5,624,675	Genital lubricants containing *zinc salts to reduce risk of HIV infection	20

Figure 8. DIALOG file 654—U.S. patents fulltext 1990−1998.

S1: 26277 (zinc or zn) and (cold or rhinovir* or antirhinovir*)

QPAT Rank	QPAT Rating	Patent No.	Title	DIALOG Rank
1.	1194	USRE33465	Method for reducing the duration of the common cold	>50
2.	1188	US4956385	Method for reducing the duration of the common cold	48
3.	1176	US5286748	General method of shortening the duration of common colds by	>50
4.	1136	US5409905	Cure for commond cold	2
5.	1100	US5622724	Spray preparation for treating symptoms of the common cold co...	>50
6.	988	US5095035	Flavor stable zinc acetate compositions for oral absorption	39
7.	985	US5002960	N-haloalkyl-4-(isoxazol-5-yl)alkoxy benzamides	>50
8.	978	US5002970	Flavor masked ionizable zinc compositions for oral absorption	12
9.	948	US5514778	Anti-picornaviral agents	>50
10.	946	US5208031	Sexual lubricants containing zinc as an anti-viral agent	13

Figure 9. QPAT-US, full text, 1990−1998.

has added the indexing from the CAPLUS database to the USPATFULL database for all of the chemical patents indexed in CA or equivalent to patents indexed in CA, so it was possible to cross the registry numbers directly into the USPATFULL file. The search retrieved seven patents, all of which had been judged by an indexer to describe one or more of the compounds in the dorzolamide family. The earliest of the patents, U.S. 4,797,413, is the patent that was extended, the one relied upon by Merck to protect the product Trusopt. A closed substructure search for dorzolamide in the MARPAT database was also performed. That search retrieved six patents with Markush structures encompassing dorzolamide. Four of those were European patents that were not equivalent to any of the U.S. patents retrieved in USPATFULL. All of these patents are relevant to a search for dorzolamide because they cover the compound itself specifically or cover closely related compounds generically.

The next test of the relevance of relevance ranking, shown in Figure 7, was to compare these unranked patents with a ranked set of patents retrieved by searching the full text of the patents in several databases. This time the databases were searched for the generic name "dorzolamide", ORed with two code numbers assigned to dorzolamide, and "Trusopt", the brand name of dorzolamide hydrochloride. The same search was done in the STN USPATFULL file, DIALOG's PATFULL file cluster, and the QPAT-US fulltext file. In each case, the search retrieved 26 patents, many more than were retrieved by searching for the chemical structure of dorzolamide or the name dorzolamide alone. Not surprisingly, the text searches did not retrieve the most relevant patent, U.S. 4,797,413—patent applications are filed too early in the development of a drug for the generic and brand names to have been assigned, while pharmaceutical companies seldom disclose the code numbers assigned to compounds under development in patents. Figure 7 shows the rankings of the patents, first showing those retrieved by searching the REGISTRY file and, MARPAT, with the remaining patents listed in order of the rankings assigned by the STN FOCUS

command. The DIALOG TARGET results and the QPAT ratings are shown in the following columns. The relevance rankings of the patents are very interesting. Only two of the patents indexed by CAS to specific or generic chemical structures, shown in boldface in the table, were retrieved by the text searches. The relevance rankings of those two patents differ significantly among the search engines. STN's FOCUS command ranked them very high, DIALOG's TARGET command ranked them in the middle, and QPAT ranked them relatively low. The patents ranked lower by FOCUS show less of a discrepancy among the search engines. Nevertheless, even though it is not possible to judge from the titles whether the patents would be relevant to a specific search that might have been intended, it is clear that some or all of these search engines have failed to identify the most relevant patents.

Different kinds of searches give different results. Figure 8 shows the result of a text search for patents covering the treatment of rhinovirus infection with zinc compounds in the latest section of the DIALOG PATFULL file, which retrieved 12 219 patents. This search inevitably retrieves a large proportion of false drop, any disclosure of metallic zinc in any invention where low temperatures are involved, for example. When these patents were analyzed with the TARGET command, only two of the top 10 patents were actually about biological uses for zinc compounds, and only one related to use against the common cold. The last column of the table shows that most of the patents in the DIALOG top 10 were not within the top 50 retrieved with the same strategy in QPAT. Figure 9 shows the result of that search in the 1990−1998 segment of QPAT, which retrieved 26 277 patents. This is a larger number than were retrieved in the DIALOG PATFULL File because QPAT uses automatic stemming as well as truncation. QPAT's top 10 rated patents are all biologically related, and seven are actually related to treatment of the common cold. Only five of the top 10 patents retrieved by QPAT were among the top 50 patents retrieved by DIALOG's TARGET command.

Not only do the relative rankings of patents change when different search engines are used for the same search but also different results are obtained when a single search engine is used to search different fields of the same patents. A search for patents covering the treatment of rhinovirus infection with zinc compounds in the IBM file was done three different ways. Figure 10 shows the patent numbers of the top 12 ranked patents obtained from searches of all fields (first-page fields, abstract, and claims) compared with the rankings of the same patents in searches of the claim text and of the abstract text. Among the top 50 patents retrieved in the search of all fields, five of the six that relate to biological uses of zinc compounds, shown in boldface, were rated within the six top-rated patents. As can be seen, of the relevant patents in the top rated group found by the full file search, only two were found by searching claim text, and all were given lower ratings in the abstract text search. Only four of the 12 patents were rated among the top 50 in all three searches. How might one suppose those patents were assigned such high ratings if the search terms responsible for the ratings were not found in the claims or abstracts? There is evidence that most of them were in the titles of cited references. In a database with only the claim and abstract text, the full titles and bibliographic details of the cited references make up a very large

(Zinc or zn) and (cold or rhinovir* or antirhinovir*)

Full File – First Page, Claims and Abstract				Claims Only		Abstract Only	
Rank	Rating	Patent No.	Title	Rank	Rating	Rank	Rating
1 (Tie)	**100%**	**5409905**	**Cure for commond cold**	**3 (Tie)**	**93%**	**1**	**88%**
1 (Tie)	**100%**	**5002970**	**Flavor masked ionizable zinc compositions for oral absorption**	**>50**		**13 (Tie)**	**78%**
1 (Tie)	**100%**	**4956385**	**Method for reducing the duration of the common cold**	**>50**		**2 (Tie)**	**82%**
1 (Tie)	**100%**	**RE33465**	**Method for reducing the duration of the common cold**	**29 (Tie)**	**86%**	**2 (Tie)**	**82%**
5	97%	4830685	Wear-resistant alloy of high permeability and method of producing the same	1	97%	>50	
6	**96%**	**5095035**	**Flavor stable zinc acetate compositions for oral absorption**	>50		**13 (Tie)**	**78%**
7 (Tie)	94%	4409036	Aluminum alloy sheet product suitable for heat exchanger fins and method	2	94%	>50	
7 (Tie)	94%	3960607	Novel aluminum alloy, continuously cast aluminum alloy shapes, method of preparing semirigid container therefrom, and container stock thus prepared	3 (Tie)	93%	>50	
9	93%	5070119	Flexible intumescent coating composition	8 (Tie)	90%	13 (Tie)	78%
10 (Tie)	92%	5462712	High strength Al-Cu-Li-Zn-Mg alloys	8 (Tie)	90%	4 (Tie)	80%
10 (Tie)	92%	4642141	Method for producing grain-oriented silicon steel sheets	5 (Tie)	91%	>50	
10 (Tie)	92%	3989548	Aluminum alloy products and methods of preparation	5 (Tie)	91%	>50	

Figure 10. IBM patent server - fielded searches.

(Zinc or zn) and (cold or rhinovir* or antirhinovir*)

USPTO Rank	USPTO Rating	Patent Number,	Title	IBM Rank, All	IBM Rank, Abstract	IBM Rank, Claims
1.	(100)	5,708,142	Tumor necrosis factor receptor-associated factors	>50	>50	>50
2.	(071)	5,498,538	Totally synthetic affinity reagents	>50	>50	>50
3.	**(060)**	**RE33,465**	**Method for reducing the duration of the common cold**	**1 (Tie)**	**2 (Tie)**	**2 (Tie)**
4 (Tie)	(051)	4,517,029	Process for the cold forming of iron and steel	>50	>50	>50
4 (Tie)	**(051)**	**4,956,385**	**Method for reducing the duration of the common cold**	**1 (Tie)**	**2 (Tie)**	**>50**
6 (Tie)	(042)	5,498,322	Aluminum alloy cathode plate for electrowinning of zinc	>50	13 (Tie)	>50
6 (Tie)	(042)	3,982,055	Method for zincating aluminum articles	>50	4 (Tie)	>50
8.	(041)	5,670,319	Assay for tumor necrosis factor receptor-associated factors	>50	>50	>50
9.	**(036)**	**5,409,905**	**Cure for commond cold**	**1 (Tie)**	**1**	**3 (Tie)**
10 (Tie)	(032)	5,625,033	Totally synthetic affinity reagents	>50	>50	>50
10 (Tie)	(032)	5,675,216	Amorphic diamond film flat field emission cathode	>50	>50	>50
13 (Tie)	(030)	3,960,607	Novel aluminum alloy, continuously cast aluminum alloy shapes, method of preparing semirigid container stock therefrom, and container stock thus prepared	7 (Tie)	>50	4 (Tie)
13 (Tie)	(030)	4,186,034	Method of manufacturing aluminum alloy sheets containing magnesium and zinc	39 (Tie)	4 (Tie)	>50
17 (Tie)	**(029)**	**5,002,970**	**Flavor masked ionizable zinc compositions for oral absorption**	**1 (Tie)**	**13 (Tie)**	**>50**
22 (Tie)	(024)	3,941,619	Process for improving the elongation of grain refined copper base alloys containing zinc and aluminum	14 (Tie)	>50	18 (Tie)
22 (Tie)	(024)	5,496,426	Aluminum alloy product having good combinations of mechanical and corrosion resistance properties and formability and process for producing such product	21 (Tie)	13 (Tie)	18 (Tie)
26 (Tie)	(023)	4,285,739	Process of manufacturing solid bodies of copper-zinc-aluminium alloys	21 (Tie)	>50	18 (Tie)
26 (Tie)	**(023)**	**4,503,070**	**Method for reducing the duration of the common cold**	**1 (Tie)**	**4 (Tie)**	**>50**
26 (Tie)	**(023)**	**5,622,724**	**Spray preparation for treating symptoms of the common cold containing unchelated ionic zinc compounds**	**>50**	**13 (Tie)**	**>50**
31.	**(021)**	**5,095,035**	**Flavor stable zinc acetate compositions for oral absorption**	**6**	**13 (Tie)**	**>50**
32 (Tie)	(019)	4,067,753	Process for the manufacture of shaped parts from multi-component silver-copper alloys	21 (Tie)	>50	18 (Tie)
44 (Tie)	(014)	5,470,403	Cold rolled steel sheet and hot dip zinc-coated cold rolled steel sheet having excellent bake hardenability, non-aging properties and formability, and process for producing same	49 (Tie)	>50	34 (Tie)
44 (Tie)	(014)	5,503,689	General purpose aluminum alloy sheet composition, method of	>50	13 (Tie)	>50
44 (Tie)	(014)	5,531,962	Cadmium-free silver alloy brazing solder, method of using said solder, and metal articles brazed with said solder	>50	13 (Tie)	>50

Figure 11. USPTO full file search compared with searches of the IBM patent server.

Set #	Expression	Results
S1	Ethylamino dihydro methyl thieno thiopyran sulfonamide	346545

Rank	Rating	Patent No.	Title
1.	1120	US4677115	Antiglaucoma thieno-thiopyran and thieno-thiepin sulfon
2.	**1115**	**US5441722**	**Short synthesis of 5,6-dihydro-(S)-4-(ethylamino)-(S)-6-[C.su...**
3.	1110	US5308863	Substituted aromatic sulfonamides as antiglaucoma agents
4.	1110	US4863922	Substituted aromatic sulfonamides as antiglaucoma agents,
5.	1110	US4824968	Substituted aromatic sulfonamides as antiglaucoma agents
6.	1110	US4820848	Substituted aromatic sulfonamides as antiglaucoma agents
7.	**1110**	**US4797413**	**Thieno thiopyran sulfonamide derivatives, pharmaceutical**
8.	1108	US5474919	Bioconversion process for the synthesis of transhydroxy sulfo...
9.	1108	US5206240	Nitrogen-containing spirocycles
10.	1108	US5175284	Tricyclic thienothiopyrans as pharmaceutical intermediates
11.	**1108**	**US5091409**	**4-alkylamino-6-(C.sub.3-5 -hydrocarbyl)thieno[2,3-B]thio-**
12.	1104	US5334591	Tricyclic thienothiopyran carbonic anhydrase inhibitors
13.	1104	US5120757	Substituted aromatic sulfonamides as antiglaucoma agents
14.	1096	US5308842	Tricyclic derivatives of the thienodiazocine and thienothiadi...
15.	**1094**	**US5620970**	**Topical ophthalmic carbonic anhydrase inhibitor formulations**
16.	1084	US5633247	Nitrogen-containing spirocycles
17.	**1064**	**US5688968**	**Enantioselective synthesis of 5,6-dihydro-(S)-4-(ethylamino)-...**
18.	1062	US5574176	Synthesis of an intermediate for the preparation of 5,6-dihyd...

Figure 12. QPAT-US full text search dorzolamide chemical name fragments.

part of the searchable text. A review of the full records for some of the patents that sounded irrelevant showed that they were rated highly because there were citations to books published by "Cold Spring Harbor Press".

When the same search is done on the USPTO database, the results are distinctly different. Figure 11 shows partial results of the same search in the USPTO database, which contains first page data and abstracts but not claims. The table shows the top 10 ranked patents from the USPTO search followed by other patents among the top 50 that were retrieved by the searches in the IBM file, with the rankings provided by searching the IBM file using the full file, the abstract text, and the claim text. Fully 20 of the top 50 patents retrieved by the USPTO search engine were biologically related, and seven were directed to the treatment of the common cold with zinc compounds. Only 18 of the 50 top ranked patents in the USPTO database were among the top 50 ranked by any of the three searches in the IBM database, while the relative rankings of patents retrieved in both databases varied significantly. Relying upon a search engine that supplies only the 50 or 200 top-rated patents is particularly unreliable, especially in a system such as IBM's that ranks thousands of patents on a scale of 100. The top 50 patents in the full file search had relevance ratings between 100% and 84%, with many tied at the same rating. The top 50 patents in the search of claim text had relevance ratings between 97% and 84%, also having many patents tied for the same rating. The relevance ratings for the top 50 patents in the search of the abstract text were clustered between 88% and 78%, with 38 patents tied for the lowest rating of 78%. It is reasonable to assume that many other patents would also share the 78% rating, Although the search in the USPTO database was also truncated after the top rated 50 patents, the range of ratings is much wider; the top-ranked patent was assigned a rating of 100, and the lowest of the 50 was assigned a rating of only 14.

Relevance ranking gives unreliable results, but there are things it can do that are impossible with old-fashioned Boolean logic. The search for dorzolamide patents was unable to retrieve the patent that claimed dorzolamide because the compound was described by its chemical name instead of the drug's generic or brand name. Searching for chemical compounds by using name fragments has always been difficult. The number of fragments and the type and placement of punctuation in patent examples is unpredictable. One can never know in advance whether the specific compound one is looking for is in an example or not and, if it is, how the nomenclature fragments are organized when the compound is named in the application. Attempts to solve the problem by segmenting the chemical names for the database index are helpful, but its still very difficult to guess what kind of proximity operators will retrieve a reference to a compound. Searching with a relevance ranking algorithm, however, presents an opportunity to retrieve the patents that disclose the compounds of interest without using proximity operators at all. In Figure 12, all of the name fragments in dorzolamide's chemical name were used as search terms in the QPAT-US full text file with the default OR operator and retrieved 346 545 patents. The relevance ranking algorithm, however, placed five of the seven patents that were retrieved in the chemical structure search, those shown in boldface, among the highest ranking 18 patents. This is an impressive result, but one could not rely on the same results for every search. This technique is, however, a remarkably cheap and effective way to search for a chemical structure when comprehensive results are not required.

Is relevance ranking relevant to patent searching? It is not comprehensive, and one cannot rely upon relevance ranking

to identify all of the relevant patents in all searches. Because relevance ranking algorithms deal only with natural language, it is useless for interpreting the nontextual information provided by drawings and chemical structure diagrams. But in some cases, relevance ranking can identify patents that focus on the concepts of interest and reduce the time required for screening search results. It cannot replace the other search techniques used in patent searching; it is simply one more tool that can be used to supplement the other techniques for searching patent databases.

REFERENCES AND NOTES

(1) Donovan, K. M.; Wilhide, B. B. A User's Experience with Searching the IFI Comprehensive Database to U.S. Chemical Patents. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 139−142.

(2) Survey of Chemical Notation Systems. National Academy of Sciences-National Research Council Publication 1150. NAS−NRC, Washington, 1964.

(3) Meyer, E. Computer representation and handling of structures: retrospect and prospects. *J. Chem. Inf. Comput. Sci.* **1991**, *31*(1), 68−75.

(4) Barnard, J. M. Online graphical searching of Markush structures in patents. *Database* **1987**, *10*(3), 27−34.

(5) The American Heritage Dictionary, 3rd ed.; Houghton Mifflin C.: Boston, MA, 1992.

(6) Feldman, S. "Just the Answers, Please": Choosing a Web Search Service. *Searcher* **1997**, *5*(5), 44−57.

(7) Brenner, E. Sorting out the competitors. *Monitor* October, **1994**, 164.

(8) Lambert, N. The idiot's guide to patent resources on the Internet. *Searcher* May 1995, *3*(5), 34-9.

(9) Lambert, N. More patents -- lots more -- on the Internet. *Searcher* Nov−Dec 1995, *3*(10), 24−27.

(10) Lambert, Nancy. But What is in It for IBM? "Free" Patents on the Net. *Searcher* Sept **1997**, *5*(8), p 33−37.

(11) Lambert, N. QPAT-US: A new patent search tool for the Internet. *Database* Aug−Sep 1996, *19*(4), 56.

CI9801421