of oxygen and moisture does get in to ensure desirable product properties. That simple film may consist of two, three, five, or seven layers. There may be core structural layers; surface heat seal layers or cling layers; oxygen and/or moisture barrier layers; adhesive tie layers to hold the thing together. There may be tackifier additives, antioxidants or prodegradants, crystallization nucleating additives, plasticizers, or optical clarity promoters. One or more layers may be a blend rather than a single polymer. That simple film is probably not simple at all, but highly complex.

Now if I am trying to find whether a novel kind of polymer, or novel kind of additive, has been used in a film, it probably will not be too difficult. But the questions my colleagues and I end up dealing with involve different combinations of those same old commodity substances, the good old simple polyolefins. The potential hits are many, the differences from the prior art subtle. It is questions such as this that make me pray fervently for the success of Derwent's plans that, if achieved, would let me discriminate in the details of a single polymer's structure, and its blends with a second polymer, and separate them from another polymer system (which could involve similar or identical polymers) present in the same overall structure. It is questions such as this that remind me over and over that discrimination can be one of the highest virtues and not just a dirty word.

So here's to discrimination in its most positive sense; may its presence in polymer databases grow. May the database producers develop better ways of implementing discrimination in their systems, and may their staffs be able to apply these improved systems in a consistent fashion that will be useful for us—because a system that breaks down repeatedly is a system that cannot be trusted, and will not be used. May the online hosts who deliver the databases develop improved methods of handling systems that can become quite complex, if they start involving multiple layers of linking as has been proposed by Derwent.

May I offer a final prayer. Maybe it could even be possible for some of the competing database producers to work out some sort of coordination of their efforts, so that work duplicated by two or more producers could be consolidated, and each could concentrate on the special features that are its particular strengths. To a significant degree Derwent and API have done this in the petroleum and petrochemical area, for almost 20 years and with considerable success, and without violating antitrust considerations. Is it totally ridiculous for me to propose a broader implementation of cooperation? Cooperation that might result in less duplication of effort and free resources to do things that might look uneconomical today? Nine years ago, at a meeting in this very same hotel in Atlanta, I made some suggestions about synergistic database combination that might benefit information users. I admitted at the time that it seemed utopian, but some of what I suggested then has actually come to pass, and more is possible in the future. Perhaps if some of the leadership of the information industry were to get its focus out of courtrooms and onto issues such as these, we might see some of this take place too—with obvious benefits for information users and great benefits for information providers as well.

REFERENCES AND NOTES

(1) Kaback, S. M. Polymer Patent Information Systems Could Be Even Better! *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 371-379.

(2) Briggs, J. A.; Ferns, E. A.; Shenton, K. E. Improvements in Derwent Plasdoc System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 454-458.

Online Searching of Polymer Patents: Precision and Recall

NANCY LAMBERT

Chevron Research and Technology Company, P.O. Box 1627, Richmond, California 94802-0627

Received June 1, 1991

Patent information specialists searching online databases for patents in polymer subjects have faced a number of problems in precision (how clean the search is) and recall (how comprehensive the search is). Some databases have been designed for high precision, but recall in these databases can be both poor and inconsistent over time. Other databases have been designed for high recall, and the searcher faces looking through many irrelevant references. This paper discusses precision and recall problems in three major databases that cover polymer patents: Chemical Abstracts, the Derwent World Patents Index, and the IFI Comprehensive Index to U.S. Chemical Patents. It mentions some solutions that the database producers are proposing and discusses additional solutions to be considered.

INTRODUCTION

One of the oldest problems in computerized information retrieval is how to balance precision and recall, and optimize both. Precision is also known as the relevance of a search: how many of the references retrieved in the search were, in fact, pertinent. Recall is the comprehensiveness of the search: how many of the relevant references actually in the database the search retrieved. Numerous case studies have attempted quantitative measurements of precision and recall in great statistical detail in all possible subject areas under all sorts of circumstances.

The reader will be relieved to know that this paper will not add to their number. Instead, it will look qualitatively at three major databases that index patents in polymer chemistry, discuss how well their indexing policies and systems work in terms of precision and recall, look at specific precision and recall problems that occur in some or all of them, and describe some solutions that might help with these problems.

The databases are

1. Chemical Abstracts, whose online file contains polymer registrations and international patents on polymer chemistry since 1967.

2. The Derwent World Patents Index, which has covered international polymer patents since 1966.

3. The IFI Comprehensive Index to U.S. Chemical Patents, which has covered U.S. polymer patents with its current indexing system since 1972, some aspects since 1964.

Improvements for all three databases are either being planned or actually under development that will clear up some of the problems discussed in this paper. But, with a few

0095-2338/91/1631-0443$02.50/0

exceptions, those solutions will be applied only from time of implementation forward. And old patents do not go away. They are still of vital importance, for instance, for patentability and validity questions. So searchers will still have to work with the old indexing systems. They will be best armed knowing just where each one's strengths and weaknesses lie.

CHEMICAL ABSTRACTS SUMMARY

Chemical Abstracts (CA) has always been designed for precision in searching, and its indexing policies to this end have been applied to polymers. Polymers, both homopolymers and copolymers or higher, are registered separately, normally in terms of their component monomers. But, under specific well-defined circumstances, the same polymers are also registered with different registry numbers by the structural repeating unit (SRU), either alone or with certain chemically modified end groups. In other circumstances, for instance when the exact starting materials are not specified, polymers are registered only by SRU. The same polymer is registered separately when made from different monomers—say, a polyester from a diacid and from the corresponding acid halide. Polymers from the same monomer but with different tacticities are registered separately. Since 1987, block, graft, and alternating copolymers have been registered separately from the same random copolymers. And so on. The searcher must keep in mind that a polymer of interest may have several different registry numbers and must know and apply CA's rules to be sure of getting all the registry numbers that apply to that polymer. To make this job easier, Chemical Abstracts Service (CAS) is planning to link these registry records for a single polymer by cross-referencing the registry numbers. However, the searcher's problems are not over yet.

Finding a complete list of applicable registry numbers can be difficult, especially if the search criteria include broad classes of chemicals. The searcher may not be able to run a broad structure search for monomers or SRUs in the registry, because of system limitations. And a search in the registry using any available search parameters—structures, component registry numbers, dictionary searching, or some combination—may retrieve too big a set of polymers to transfer into CA. For instance, a search by component registry number of styrene-containing polymers produces almost 28 000 separate registry numbers.

Suppose that the searcher overcomes these hurdles, generates a complete list of applicable registry numbers, and transfers them from the registry into CA. Even then, retrieval is not complete. Registry numbers will retrieve only patents in which those compounds are spelled out in examples and preparations or elsewhere in the patent specification. But patents all too frequently talk about and claim chemicals in terms too generic for registry numbers to apply. For example, polymers and their monomers may be described in Markush structures, with variable groups. CAS normally will not generate specific chemicals from these Markush structures and apply the appropriate registry numbers, although in some cases a base registry number may be indexed as "derivative". Nor is CAS's Marpat, which addresses Markush structures, being applied to polymers. Or, polymers may be described in a patent in terms of "laundry lists"—copolymers with lists of possible components for each monomer. Again, CAS will not generate the specific copolymers and index the registry numbers. To find patents in which the desired polymers are not listed specifically but are included in these broader cases, the searcher must depend on polymer class terms and other terminology searching. Also, before 1987, the most common polymers were not indexed with registry numbers at all when they appeared in the literature in fiber or rubber applications. This policy is a holdover from the CAS printed index, in which the searcher, looking up the CA index name for, e.g., polycaprolactam in the Index Guide, finds the cross-reference, "fibers—see *polyamide fibers*". No such cross references exist in CA online, of course.

CAS does not always index polymerization catalysts. If the use of the catalyst is incidental to the main point of the patent, the catalyst is not indexed. Very common constituents of multicomponent catalysts are usually not indexed. To search them, one must depend on their being mentioned in the abstract—which means, of course, that one must search CA on a host that includes abstracts.

In summary: Chemical Abstracts will usually give a precise search; but, especially for patents, it cannot be depended on for comprehensiveness.

DERWENT SUMMARY

The Derwent World Patents Index lies at the other extreme: Its polymer coding system was designed back in 1966 for highest possible recall—a situation that causes far more serious precision problems now, when Derwent includes about 900 000 polymer patents, than it did in the late 1960s.

Some background: Derwent has divided the world of technology into subject sections. Sections A–M cover aspects of chemistry; P is general; Q is mechanical; and S—X cover electrical sections. Derwent then applies multiple levels of indexing, the most in-depth of which are used in some but not all of these subject sections. Polymer chemistry is Section A; the Derwent polymer coding is applied only to patents in Section A. Derwent's Markush DARC structure indexing and its predecessor, the chemical fragmentation codes, are applied only to the pharmaceutical, agrichemical, and general chemical sections (B, C, and E).

As mentioned, Derwent's polymer codes emphasize depth of coverage. They are hierarchical; the broader codes are also posted when the narrower codes are indexed. For instance, searching the broad code for aliphatic and cycloaliphatic olefins retrieves all the specifics—ethylene, propylene, and so on. The searcher need not worry about standard vs actual monomer forms. The general code for "terephthalic" is applied as well as separate codes for acids, anhydrides, acid halides, and so on; the searcher may choose at which level to search. Derwent indexes the claims and examples and adds significant information from the body of the patent. It indexes all applications mentioned and all properties that indicate improvements over the prior art. And it indexes Markush structures and laundry lists as well as specific compounds.

However, the Derwent polymer coding was *not* designed for precision. Modifications in 1968, 1972, 1977, and 1982 added specific search terms but did not change the basic concept. In the original system, each subconcept—broad or narrow—has a separate code; and strings of codes linked together represent whole concepts. However, in a patent record these codes are linked at only one level, and that a very broad one; so searching linked codes produces vast numbers of false hits. The key serial system introduced in 1978 sacrificed some of the breadth of retrieval for increased precision by prelinking the strings of separate codes indexed for specific concepts. However, the resulting rigidity was so limiting that Derwent is going back to the old code concept with its soon-to-be-introduced new polymer indexing.

Derwent's main problem with precision is its very limited number of codes for specific chemicals. Derwent provides codes only for very common specific monomers (and key serials for a few specific copolymers since 1982) and lumps other monomers into generic "other" codes in the appropriate sections. Even fewer codes are available for specific chemical-modifying agents and polymerization catalysts. A polymer chemical registry introduced in 1984 contained only about 750 compounds.

The proposed new polymer coding system will partially alleviate this problem. It increases the number of specific chemicals indexed to several thousand. Chemicals not included will be indexed with very broad "chemical aspects" designed for the new polymer coding system. However, Derwent has never implemented its users' most urgent suggestions over the years that chemicals indexed only with "other" codes in the polymer section also be structure-indexed with chemical fragmentation codes from the chemicals section. Nor does Derwent plan to use either Markush DARC or full fragmentation indexing on polymer chemicals not indexed specifically in the new system.

Derwent's polymer coding does have one recall problem: A very significant number of patents on polymers, especially applications, are not classified into the polymer section at all; so the polymer coding is never applied. Some examples: patents mentioning polyester, nylon, or elastomers *in their titles*, of which 1%, 7%, and 36%, respectively, are not in the polymer section.

IFI SUMMARY

The IFI Comprehensive Index strikes a middle ground between CA and Derwent: Its indexing is not so precise as CAS registry numbers nor yet so broad as Derwent's polymer coding. IFI, like Derwent, indexes generic structures and laundry lists as well as specific polymers. It too indexes polymers both in terms of their monomers and in broad polymer classes. In fact, IFI's "collection groups" of indexing terms allow very broad class retrieval indeed: One can search, for instance, for all addition polymers or all condensation polymers.

IFI has its own smaller version of a chemical registry. Monomers, modifying agents, polymerization catalysts, and so on which are included in their thesaurus of about 14000 chemicals are indexed as specific chemicals; other less common compounds or compounds represented in patents with Markush structures are indexed with IFI's chemical fragmentation system. Monomers and modifying agents, both specific compounds and fragments, are linked to the polymer roles described in the paper by Monica Rieder.[1]

IFI has decreed standard starting materials for each of its seven broad classes of polymers and indexes these in the appropriate polymer roles whatever the actual starting material the patent mentions. So the searcher need not worry about searching the acid chloride to find a certain polyester; the diacid gets everything. In addition, IFI indexes the actual starting materials if the patent claims preparation processes. IFI also has fairly extensive vocabulary of specific homopolymers and copolymers. If this vocabulary includes the polymers wanted, retrieval is very precise.

IFI, by definition, includes only U.S. patents—one major problem in recall. IFI proposed several years ago a cooperative venture with Derwent in which IFI indexers would apply IFI polymer indexing to documentation abstracts of international patents in the Derwent polymer section. Unfortunately, this proposal did not receive enough support to go forward.

SPECIFIC PROBLEMS

What kinds of problems do searchers run into with these systems? Lots of them. What are possible solutions? They vary.

1. A problem arises when what is of interest is not the polymer chain itself but what is attached to that chain in a posttreatment. With some exceptions, CAS tends not to register modified polymers separately, but rather to index them with the registry number of the parent polymer and add text indexing in CA for the modifications. The searcher must then think of all the terminology that could be used to describe the modification or modifying agent, or come up with registry numbers for all the specific chemicals that fit into the class of modifying agents of interest—an impossible task when this is a broad class of chemicals.

Table I

component 1	component 2
ethylene	acrylic acid
propylene	ethyl acrylate
	methyl acrylate
	methacrylic acid
	methyl methacrylate
	ethyl methacrylate

Derwent, as mentioned, codes all but the very common modifications only with "modified polymer—other" codes. The new polymer coding system should index modifying agents more precisely and also link them to some sort of "modifier" tag term.

IFI's practice, that of compound- or structure-indexing the modifying agent in the appropriate polymer roles, works fairly well. However, IFI has precision problems if the modifying agent could also be a monomer—for instance, maleic anhydride. A copolymer modified with maleic anhydride and an otherwise-modified terpolymer of maleic anhydride would be indexed the same way. IFI also has serious precision problems if the modifying agent is not indexed as a specific chemical and has a chemical fragmentation similar to the monomers making up the polymer chain. One possible solution in IFI is an enhancement to the role system so that modifying agents are distinguished from monomers.

2. For block copolymers, none of the databases currently specify the order of the blocks (AB, ABA, etc.). For graft copolymers, none of the databases specify which monomer is the base chain and which is the grafted entity. Such distinctions can and should be incoporated into any polymer indexing system, possibly by linking roles or descriptive terms to the monomers.

3. Oligomers are a problem in all three databases, as anyone knows who has ever had to search for surfactants in any of their multitude of applications. The first questions that always arises is: It is a polymer or isn't it? All three databases have rules, based primarily on numbers of repeating units, which decree when a chemical becomes a polymer and is indexed as such. CAS's rules take into consideration whether the oligomer is an exact structure, applying a separate registry number when it is. If the chemical of interest falls into the grey area—say, a compound with between 3 and 15 repeating units—then it must usually be searched both as a polymer and as a nonpolymeric chemical. Searching it as a nonpolymeric chemical can lead to further problems, especially if the repeating unit is from an olefin. The chain then becomes an alkyl tail: unsearchable in IFI; very imprecisely searchable in Derwent; and frequently unregistered in CA if its structure is inexact. Whether or not the compound qualifies as a polymer, all three databases should index the number of repeating units in an oligomer struc-

ture. This is necessary whether the oligomer is the whole compound or merely a tail attached to a base structure.

4. All three databases need to work with the online hosts to index and make searchable all sorts of number-range information when this information is specified in the patent—not only the number of units in an oligomer, as mentioned above, but also monomer ratios in copolymers; polymer ratios in blends; molecular weight ranges, at the very least by low-medium-high-ultrahigh; ranges for numbers of modifying groups attached to a polymer chain; temperature and pressure ranges for processing steps; ranges for numbers of atoms in Markush groups; etc.

5. None of the three databases distinguishes specific polymers actually exemplified from those included in the broadest claims of the patent, the Markush structures and laundry lists mentioned before. All three databases should make this distinction with major-term/minor-term indexing of polymers, so that the searcher can choose the level of precision wanted.

6. All three databases also need to make better use of the linking capabilities of the online hosts, to increase precision. Most hosts have linking available at multiple levels—subfield, field, sentence, paragraph, whatever; the names vary with the hosts—but all three databases link polymer indexing at most at one level. CA links polymer registry numbers and/or terms to text modifications which can cover a host of concepts. IFI links fragments to each other within compounds, and it links chemical indexing, whether by specific chemicals or fragmentation, to roles. But IFI does not currently link monomers within a copolymer. Derwent uses one level of linking with varying rules for when separate link sets are applied. Most of these are at a fairly macro level—for instance, the polymer used in fabrication equipment is in a separate link set from the polymer being processed with that equipment.

Derwent is proposing a three-level linking scheme, described in the paper by Julie Briggs.[2] A somewhat different vision of three-level linking: linking the chemical components of a monomer; linking the monomers within a copolymer; linking polymers and copolymers within a blend or construct. At all levels, the polymers can and definitely should be linked with any appropriate tag terms such as "blend" or "laminate" and also with their applications—something that IFI and Derwent do not now do.

7. Patents will frequently claim multicomponent systems in a laundry list form in which multiple options are listed for each component. IFI and Derwent both have precision problems when the searcher is looking for a multicomponent system: A search will retrieve patents indexed for the desired components as either/or options of the same component, in another multicomponent system. For instance, suppose that a patent claims copolymers of either ethylene or propylene with acrylic or methacrylic acid or a variety of alkyl acrylates and methacrylates (see Table I). IFI and Derwent index all the appropriate monomers separately, linking them to the appropriate copolymer codes in Derwent or roles in IFI. But suppose that the searcher wants patents on an acrylic acid–ethyl methacrylate copolymer and searches in IFI or Derwent using the monomers (since this copolymer is not indexed specifically in either IFI or Derwent). The search will retrieve this patent, even though it does not cover acrylic acid–ethyl methacrylate copolymers, but merely lists acrylic acid and ethyl methacrylate as two options for component 2.

A solution to this problem has existed since late 1989 at the American Petroleum Institute's Central Abstracting and Information Services. Using a concept that Elliott Linder originated, John Lucey developed PC-based software that performs what API calls template indexing. This software produces, on command, a template in which the indexer need only enter the appropriate components in the appropriate lines—say, one line for each of the sets of possible components and one line for applications. The computer then generates all possible permutations of one item each from each of the lines and makes each permutation into a separate link set. So someone searching for acrylic acid–ethyl methacrylate copolymer patents will no longer retrieve this patent, because none of the link sets will contain both acrylic acid and ethyl methacrylate.

API developed the template indexing system for catalyst compositions but often uses it for other applications. Predictably, API has found that the larger laundry lists generate permutations that go on for hundreds of pages. The online hosts are then faced with the problem of storing these monstrous records. IFI is looking into a variation of template indexing and has found that the hosts will indeed not be able to cope with the size of records that many polymer patents generate. So the solution, which would require a great deal of work on the part of the hosts (not to mention encouragement on the part of the users), might be for the hosts to modify their online software to store and make searchable, not all the separate permutations, but the templates themselves.

CONCLUSION

Producers of all three databases discussed in this paper are aware of precision and recall problems related to the searching of polymer patents; and all of them are working on solutions. But this immediately creates two problems. The first, as mentioned earlier, is that older patents will still be indexed in the old ways and must still be searched with all the lack of precision and recall inherent in the old systems. The other problem is that the more new indexing systems each database introduces over time, the more complicated the search process becomes for the searcher, who must know and apply all the indexing systems for all the databases, with the appropriate time ranges, to do the best possible search. This becomes a truly daunting situation for someone just entering the field. So a final wish is that, while the database producers are coming up with indexing systems that really work, the computer geniuses would also develop a magic wand to convert the tattered rags of old indexing to Cinderella's shining ball gown of the new systems.

REFERENCES AND NOTES

(1) Reider, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 458–462.
(2) Briggs, J. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 454–458.