Statistical analysis of patents applied to a separation process

A. Girard and M. Moureau
Institut Français du Pétrole, Rueil Malmaison, France

The statistical use of patents to determine the trend of research fields in an industry, country or firm is based on the postulate that there is a correlation between research and patents, and that patents can be considered as a valid indicator of technology. The truth, however, may disprove this postulate. A company may have gained such an advance in a given field that it prefers to keep its knowledge secret rather than risking, by applying for patents, to reveal a significant portion to its leading competitors. A company may, on the other hand, apply first for patents to cover technologies and products that it intends to promote, and then for ‘smoke-screen’ patents to camouflage its true choices. As such the survey analyst must crosscheck information coming from different sources so as to give the data acquired their true position.

1. Counting methods

Database producers as well as the host computers containing these databases are preoccupied with supplying users so desiring with systems enabling them to use these databases for classification and counting purposes. The ZOOM command on ESA-IRS, the GET command on Orbit or the MEMSORT command on Questel are examples of such systems. They make possible quick searches of a subject that can show, for example, who the leading patent appliers are, classified by decreasing order of the number of patents they have taken out.

Derwent has developed a special program for the statistical analysis of its patents. It is called PATSTAT PLUS and has been used with good results by information specialists at IFP. Some 20 surveys were made between March 1985 and December 1989, concerning subjects that were varied enough in terms both of structure and field to gain a good understanding of the possibilities and limitations of this type of approach.

The Derwent database is particularly well suited for counting operations because of both the scope of the patents and the way the data are entered. In terms of scope, the patents taken out in industrialised countries are covered thoroughly, at least for recent years. In terms of entering data, there is no redundancy of information in relation to the number of patents or patent applications stemming from an invention (in the ‘patent’ sense). Because of the structure of Derwent recordings, such data are entered only once no matter
how many patents were spawned. This is a necessary condition for a valid statistical study, i.e. not having multiple entries for the same invention.

Likewise, the name of organisations is re-indexed with a four-letter code, making counting easier by giving a single form to a heading that may either have varied in time or have encompassed different companies in a group under the same heading.

For example, UNVO encompasses such varied spellings as:

- UNIVERSAL OIL PROD CO
- UNIVERSAL OIL PRODS CO
- UNIVERSAL OIL PRODUCTS
- UNIVERSAL OIL PRODUCTS CO
- UOP INC.

and ESSO encompasses:

- ESSO
- ESSO RES AND ENG CO
- ESSO RES ENG CO
- ESSO RESEARCH & ENG CO
- ESSO RESEARCH & ENGINEERI
- ESSO RESEARCH AND ENGINEE
- EXXON RES & ENG CO

The one exception to this rule is the Japanese conglomerates that are so large that the companies making them up keep their own identities. For example, Mitsubishi jidosha (automotive) will not be grouped with Mitsubishi denki (electricity). They will be assigned different codes.

The procedure used consists in downloading all the references applying to a question into a microcomputer of the IBM PC-XT type. These references will then be processed in the different fields available. Counting will deal with all the elements that can validly be taken into consideration.

The number of references is closely linked to the type of question and the way it is stated. The problem lies in seeing that the question suits the way the patents have been indexed. This indexing is based on international patent classification, and on Derwent's re-indexing by 'manual codes.' Some questions will be relatively easy to ask since they will effectively be delimited by their classification codes. Others, however, will be extremely difficult, and several phases will have to be undertaken to make sure that there is neither too much silence (references missing) nor too much noise (non-pertinent references).

The number of references processed will thus vary considerably, often being in the vicinity of 1000 references.

2. Choice of a strategy

A specific example will be used to describe the methodology used. It concerns a study of separation techniques by adsorption on zeolites and molecular sieves used by companies.