Information mining and data analysis, a global tool for local action: finding a needle in a haystack.

Christian Dutheuil* , Jeanet Ginestet**, Giovanna F Miranda***

* Sanofi-Synthélabo Recherche, Paris, France

** Sanofi-Synthélabo Recherche, Montpellier, France

*** Sanofi-Synthélabo Spa, Research Centre Sanofi Midi, Milan, Italy

Information volume explodes when one calls on the resources of Internet. Engines and agents (free tools) retrieve crude full-text information that is not structured, often with a very poor level of real substance and a lot of waste, to say nothing of the risk of silence. Most of the time interesting information is in the invisible net, such as full-text articles from periodicals.

In the 1980's, bibliometric methods introduced data mining to provide information on the structure and organization of a domain. This approach uses bibliographic databases to build a corpus of information that is globally analysed by statistical or data analysis tools. An information mining station is built as a global structured puzzle of tools, but it can boost a local library service's performance to help end-users obtain strategic information that they are unable to retrieve by themselves.

The most valuable feature is structured access to information as soon as it is electronically published. The gap between the easy availability of crude data (full-text) and indexed data (databases) is filled by a new customized product that links library services to documentation offers. In the pharmaceutical industry, this new technology may be the answer the researchers' increasing needs for customized information.

This scenario makes it indispensable to change the management strategy of library services, establishing new aims, skills and objectives for every member of the team. Information professionals are used to bibliometrics and already have all the skills needed to manage and use these recent tools

Introduction

The middle of the 19th century through to the start of the 20th century saw the birth of large bibliographic databases. Automatic documentation techniques started to catch on in the 1970s as a means of managing retrieval as the output of scientific publications grew exponentially. This was the era of data bases that retrieved and indexed the content of articles in a specific structured language (Medline, Chemical Abstracts), without access to the full text.

As we all know, in the last decade the volume of information has exploded. Internet is fast becoming the preferred medium for information transfer and end users have to track progress of new items regularly - even a daily search would be justified !

Users want immediate access to articles as soon as they are published electronically, such as in recent issues of periodicals and archive sites for older publications. Search platforms have therefore been developed by publishers (Science Direct, set up by Elsevier) and by subscription agencies (Ebsco OnLine), which become information servers just like data bank servers.

Publications that are immediately accessible – summaries and abstracts, full texts, bibliographic links – are sources of instant information, often completed by advance information on articles submitted for publication: ASAP articles from the American Chemical Society; Paper in Press Alert from the Journal of Biological Chemistry ; Early Edition Contents Alert from the Proceedings of the National Academy of

Science, and summaries of future issues – for instance, the Future Table of Contents Alert from AJP Heart & Circulation Physiology.

Nowadays, librarians and information professionals are under pressure to produce less structured information, presented more informally, more like what is offered on Web sites. The Internet has provided users with continuously growing amounts of information, almost all of it presented in free, natural, unstructured language.

Engines, meta-engines (free tools) and classic agents retrieve crude information more or less as it comes, selected using largely obscure criteria, with a lot of waste. Every item of information must be assessed singly and its source, indicated by the hypertext link, must be evaluated.

All these methods of retrieval, like traditional documentation, permit only linear exploitation of the information. Bibliographic reviews try to single out the main features, but they depend on the experience and subjective approach of whoever prepares them.

We are drowning in information and do not know how to make sense of it. The human mind's capacity for compilation is rapidly saturated as the volume of documentation swells, and users' needs have led to the development of high-performing but highly specialized tools.

Bibliometrics

Early in the 20th century, the distribution of articles and of words used in texts was defined by three primary laws intended to describe the workings of the system in mathematical terms:

Zip's law essentially predicts the phenomenon that *as we write, we use familiar words with high frequency.* Lotka's law states that in *a well-defined subject field over a given period of time … a few authors are prolific and account for a relatively large*

percent of the publications. Bradford's law states that *publications have a core and scatter phenomenon, a few core journals being prolific in publishing articles while others publish progressively fewer articles.*

These laws do not provide any information on the content by themselves but opened the way to operational methods of bibliometrics that can produce useful quantitative data and analysis to understand and improve the organisation of information. Bibliometric methods for exploiting information were introduced in the 1980s, with the aim of overcoming the problems of capacity and the subjectivity of the summaries. These are now known as data mining tools. Various methods are used, based on counts of the information items found in documentary fields :

co-occurrences mapping

clustering methods

classifications

multivariate analysis, especially correspondence analysis (AFC)

relational analysis

analysis and classification by artificial neural networks

analysis using genetic algorithms.

Each approach has its own properties and limits. It is always advisable to use at least two so as to exploit their complementary features, corroborating the results and defining the limits of reasonable interpretation when establishing the structure of the domain examined :

- systemic analysis
    . proximity groups based on overall relations between all the items of information;
    . relative position of an item (concept, periodical, product, company);

- analysis of the time course of the relations between all the items;

- detection of atypical behavior and emerging events (weak signals).

The pertinence and completeness of the texts are obviously decisive features.

Original data need treatment: hierarchical condensation of descriptors, consolidation of data, use of canonical forms such as abbreviations for the names of journals, transliteration of foreign alphabets and accents. The linguistic approach is essential, completing the information professional's skills.

Since 1983 data on patents has been dealt with in the chemical and pharmaceutical domains, resulting in operational studies in our units since 1992.

Text Mining

As vast quantities of on-line text become available, there is an increasing need for some system that automatically analyzes the conceptual content of natural-language texts. These may express a vast range of information, but it is encoded in a form that is difficult to decipher automatically. Natural language processing (NLP) methods are a break-away from linguistics and have developed independently. These analyse the content of a text – or a group of texts – on the basis either of semantics (meaning the terminology and hierarchical relations between expressions), or of syntax (meaning the grammatical function of expressions), with the aim of translating, summarizing, indexing or classifying. Language analysis employs reference tools with certain defined terms (logical assumptions and nominal groups), synonyms, declensions, and if possible their hierarchical relations (all part of the semantic network) and information on the context to help eliminate ambiguity. Several different approaches have been developed:

. by general, scientific and technical dictionaries and glossaries dealing specifically with the field in question. These are combined by the linguistic engine dealing with the text (Témis, IBM, Xerox);

. by classification plans, or thesaurus designs (Acetic);

. by knowledge bases (Arisem) in which information is assembled by concept (lexicalisation, declension) linked in a semantic network. This is a highly effective approach for multilingual work, as it is easier to pair off the translations of concepts where each one is independently analysed lexically in its own language, than to superimpose corresponding semantic networks.

These reference tools are essential for the two basic operations of parsing – syntactic analysis to establish the function of the term in its immediate context – and semantic analysis – to analyse the meaning of the content –. Relations can be established between concepts in order to classify documents on the basis of their content. An important corpus of documentation could be built up, linking the different components of a subject. Language analysis is one method of classification. Then there is the cluster method, classification by neuronal networks and learning techniques, and pattern matching.

Europe, and France in particular, have pioneered application in the field of documentation (IBM's ECAM, Langage Naturel Ingénia, SAPD, ERLI) and coupled it with data-mining. Simple representation tools such as co-occurrences, for instance Trivium's U-Map or Semio's SemioMap, or the mapping proposed by certain engines (Kartoo) are not text-mining tools but aided-reading tools. Interpreting these maps is not straightforward at all, in contrast with the aims of these tools.

Information Mining

Classification can be used to index a document, opening up the field of data mining for original, non-structured information. Recently some large data-mining software publishers have acquired language analysis companies (Lexiquest with SPSS software) so to as put together an information mining tool and complete the processing chain: text mining + classification tagging + data mining.

In addition to information mining, classification based on company taxonomy can also be used in a documentation portal to build up customized interfaces. Once the user's needs have been defined, the specific information products surface automatically (automatic plug-ins). The content of the portal grows dynamically and users have immediate access on the basis of simple notices, with no need for the traditional intermediaries of documentary products, such as profiles or alerts.

The idea of using these tools on advance information is attractive, and it would be technically feasible but there are legal hurdles. Contracts for electronic access to journals do not allow the use of robots. Attempts at storming Web servers are fast detected and the user's IP is put on a blacklist. We are still working to test how summaries and advance information can be processed, comparing the various tools so as to achieve an effective operational method that will be available as soon as software and journal publishers finish negotiating and come to the inevitable agreement.

Conclusion

Information mining may be used to create new information by exploring proximity of concept. It can merely summarize it, making the data more readable, discovering meaningful patterns and rules, and extracting hidden predictive information, or lead to hitherto unknown information.

We have tested most of these tools but have not found one that can be considered universal. This holds for language analysis and classification, and for data mining. Every publisher offers an approach with certain advantages, but defects and limits are inevitable too. They need to be collected in a *toolbox* so the user has his own range of techniques for customizing a method to satisfy his specific needs. For the time being users must be content with assembling their tools, varying in number and variety in relation to their needs and resources, in a personal information mining station.

In the pharmaceutical industry, this new technology may be the answer to researchers' increasing need for customized information. This scenario makes it indispensable to change the management strategy of library services, establishing new aims, skills and objectives for every member of the team. Information professionals, used to bibliometrics, must now acquire the information mining skills needed to manage and use these recent tools