The management of scientific information, or, how to cope with the flood

Nicolaas J.I. Mars

University of Twente
THE MANAGEMENT OF SCIENTIFIC INFORMATION, OR,
HOW TO COPE WITH THE FLOOD

Mars, Nicolaas J.I.
University of Twente, The Netherlands

Introduction

It is a platitude to note that the amount of scientific and technical information has been growing dramatically since the Middle Ages\(^1\), \(^2\). The number of scientists may have grown even more rapidly. Latin is no longer the Lingua Franca of science, and German and English are no longer the universal languages of engineering. These factors make it ever more difficult for a scientist or an engineer to find the information relevant to the job at hand. As no human being can have an overview of all the scientific and technical literature available, computer-supported ways to manage the flood of information would be of great potential benefit.

Traditionally, librarians (in the broad sense) have contributed to solving the problem by improving access methods. Better access means finding references to relevant literature as well as access to the documents themselves. In both areas, rapid progress has been achieved, but the growth of information seems to absorb immediately any progress achieved. Furthermore, easier access increases the "information overflow" of the consumers of the information. This becomes quite evident in the case of the Internet, where most users, after an initial period of fascination with all the information only some keystrokes away, give up in desperation once they have tried to answer a specific question using the Internet.

Research in the Knowledge-Based Systems Group of the University of Twente is aimed at reducing the information overload, at several levels. One approach is to support indexing. Assigning content descriptors to documents is the traditional way of
providing ways to find relevant documents. As the process of assigning descriptors is time-consuming and costly, computer-support may provide a cheaper and quicker route; whether the quality can also be improved through automation is an open issue.

Our second approach is much more ambitious, and much more risky. Rather than providing descriptors, indicating what a document is about, this project is aimed at having a computer program determine what the document has to say.

We will look at the approaches followed in the next two sections.

**Automated indexing**

In the Sapiens-project, carried out in 1990-92 for Elsevier Science Publishers, and in the recently started Condorcet-project, semi-automatic indexing methods have been, and are being, developed for scientific and technical journal articles. In both cases, assigning content descriptors is done on the basis of the title and abstract of the articles only. The reason for this is that these two items are frequently available in machine-readable format.

We believe two features of our approach are essential to providing high-quality content descriptors. The first is the use of a computational-linguistic analysis of the title and abstract of the documents to be indexed. The second is the use of a language-independent, well-structured set of concepts; in the jargon of knowledge-based systems such a set is nowadays referred to as an ontology.

So far, we have developed ontologies (in the technical sense) for a subset of medicine, for engineering ceramics, and for measurement procedures \(^2\),\(^3\). In all cases, we found that a relatively small set of atomic concepts, and an even smaller set of construction rules, enabled us to denote concisely all concepts for which index terms are to be generated.
A conceptual organization of a domain, as provided by an ontology, helps in organizing a thesaurus, as the set of terms (and phrases) corresponding to the concepts. As the ontology is independent of natural language, it can serve as the basis for thesauri in as many natural languages as desired.

We have encountered an interesting synergy in this approach. The conceptual organization of the domain, embodied in the ontology, contains a lot of highly specific knowledge about the domain. (An example would be the fact that only "body parts" can occur as the location of diseases in a medical ontology.) Such knowledge helps to clarify the ambiguousness of many natural-language expressions in the title and abstract, which would be very difficult on the basis of general syntactic knowledge only.

**Direct access to information**

Providing references to documents serves to increase, rather than decrease, the information overload. Therefore, in our Plinius-project we have set ourselves the even more ambitious goal of providing a machine-usable representation of the knowledge in scientific and technical journal articles, through semi-automatic analysis of titles and abstracts.

To give a simple example of the potential benefit of such an enterprise: representing in a computer-accessible format the information in several (preferably all) articles on experimental procedures to determine the melting point of a certain substance, would allow a computer, given sufficient background knowledge on integrating data from experiments, to provide automatically the best or consensus value of the melting point. Such integrating procedures are nowadays being carried out by standards committees (in medicine, the procedure is related to what is called meta-analysis of experimental observations). Thus, rather than boring the researchers with the flood of individual
measurements, the computer could provide him or her with the best value of the property.

We are currently experimenting with this approach in the domain of engineering ceramics. A multi-step analysis procedure is being implemented, consisting of morphological, syntactic, semantic, discourse and integration levels. (Discourse refers to the integration of information within one text; integration in our approach refers to the integration of information from several texts, as in the example of melting points given above).

Conclusion

Many problems make the realization of the vision given above a distant dream: technical, organizational, philosophical and legal problems. However, it is our strong belief that building such intelligent tools for accessing and refining scientific information are the only way to cope with the flood of scientific information.

Acknowledgment

It is a pleasure to thank Paul E. van der Vet, project leader of our Plinius project, for his comments on this article.

References


