Library Document Analysis Experiences for Comprehensive Search of the Web

Prof. Gyorgy SEBESTYEN, Ph.D., C.Sc.
Head of the Department of Library and Information Science
Head of the Ph.D. Academy of Library and Information Science
Roland Eotvos University
Muzeum krt. 608.
Budapest, Hungary, H-1088
Tel. / Fax: (36 1) 26 67 946, E-mail: lion@ludens.elte.hu

Abstract

The application of some elements of the leading universal classification systems as well as the potential uses of the search tools in professional online information retrieval in the development of a user friendly, comprehensive but information pin-pointing scheme on the Web. Analogies and differences between the Web and the big paid-for database providers are also analyzed. Special attention is paid to the use of thesauri in order to provide high performance and reliable retrieval tools on the Web.

As a researcher of the information retrieval tools in textual bibliographical databases, I consider myself belonging to the document and content analysis profession and I wish to share my experiences with the Web communities. My initiative is not an isolated action because there are currently considerable initiatives to structure the Web sources by means of classification schemes[1].

What is the Web like from the viewpoint of a library document analyst? First of all, I could compare the Web to some special aspects of the big paid-for online databases.

In fact, there are giant *multinational database providers*, such as *the Dialog Corporation*, who have several hundreds of databases so that each database covers a separate, specific domain, therefore each database is searched separately and specifically. But everything is becoming interdisciplinary, so they invented the so called *multiple search,* which means that one can make queries in up to 60 databases at a time e.g. in the Dialog.

Now why is the multiple search similar to the Web? The Web can be regarded as a giant information provider having also hundreds of various databases, plus millions of Web-sites, millions of Web documents on it and when we search for relevant information, there is no limit - no limit of 60 databases or 600 databases or more - and this fact is the cause of our first problem to be tackled.

Let us compare now the information seeking process of the big database providers with that of the Web. When using a big paid-for database, the user applies from his desktop a standardized and sophisticated *command language* which means that this language is common to all several hundred databases, but requires great competence, skill and practice. In other words, you must be a professional to interrogate these databases efficiently. Each database is *indexed by its own thesaurus,* specially developed to cover its specific domain, plus names (authors, named persons, corporate sources), space and time coordinates and other features are also specified. As a result, about 20-28 various indexed fields become retrievable, separately or combined. Boolean and proximity operators are used to combine search terms, data and expressions so as to formulate the most efficient *search strategy*. *Search results* can be displayed in standardized or user-defined formats, ranked by hits or sorted alphabetically, chronologically, etc., there are facilities for duplicate detection and elimination.

Despite some general analogies, the Web subject access tools are necessarily different. The user at the desktop opens up his *Web browser* and accesses a *search engine* or some *Web directories* - we do not deal here with navigating, surfing, etc. Big advantage: the user is not obliged to make sophisticated queries, and in most cases ranking and display-formatting is automatically made for him. At the end of this process Mr. Average will not raise questions, but when the online database specialist accesses the Web in the same way as the ordinary user, he will certainly regret not to know exactly what is happening "behind the facade".

For this reason, *in the eye of the professional online database user* a number of facts need to be clarified:

- Is it the user who actually searches the Web or just the search engine does it for him, and if it does, completely of partly?
- Consequently, does the user really search the Web, or he searches only a database of indexed Web sites?
- If the search engine is a database of indexed Web sites, is it able to cover the entire Web or just a fraction of it?
- If it covers only a fraction, what percentage of the Web is it exactly, what are the proportions of the subject fields processed, how fast can the update follow the "death and birth" of Web documents?

These questions are not raised by chance. Either for *global knowledge managemen*t or for *information pinpointing* the answers are crucial. We must be sure that

- we have a means of reviewing *all* the Web,
- we have a means of identifying every needed subject areas,
- within each subject areas we are capable of pinpointing the relevant items of information.

In this logic, the whole Web is seen like a big database provider, the subject areas identified within it correspond to the various databases in which we need to retrieve the relevant items of information. All we need is to find the appropriate tools to cope with this three-level task.

The first two levels are very closely connected. In fact, it is impossible to overview an entity without considering its sub-entities. So we need a tool that conceives the Web as a kind of universe and at the same time this tool is able to classify all the components of this universe. Why do not we use a so-called universal classification system? They are well-known and widespread wherever library services are being run.

The leading library classification systems are as follows:

- *Library of Congress Classification (LCC),*
- *Dewey Decimal Classification (DCC),*
- *Universal Decimal Classification (UDC*).

It is important to emphasize that the use of these schemes is far from being uniform. In North America the LCC is regarded as the most common system in the academic world, while the DDC is considered the most widespread in public and school libraries. By contrast, the UDC has a leading position in Europe. In conclusion, there are three different, and not one uniform world classification system and the compatibility among the three leading systems is still not resolved. That is a fundamental disadvantage.

The leading library classification systems pay a high price for their comprehensiveness, in fact, all the concept definitions and all the relationships among the components are unalterably predefined. Practically speaking that is to say that *concepts and terms* mean strictly what the classification system allows them, they *cannot be updated* as fast as to keep up with the ever-evolving Web environment, thus these gigantic universal classification schemes are cumbersome. Furthermore, the application of these systems requires sophisticated knowledge, *complicated manipulations* and painstaking intellectual efforts, which is not an advantage with end users whose majority are not library professionals.

In fact, classification has always been a laborious task. Library students take years to acquire the basic skills and after graduation, if they specialize in the classification work, they will have to toil for years to become reliable professionals. It is unthinkable to adopt for Mr. Average, the ordinary Web user, such a competence-demanding and painstaking process. But highly qualified scientists or businessmen are also unlikely to bother with lengthy searches, for the simple reason that time has a premium value in today's business environment.

We must never forget that the end user wants to get straight to the point. That means that the ideal scheme should provide him with a user friendly interface absolutely simple and absolutely easy to use.

We must also keep in mind that *the end user is a consumer*, he wants his information promptly, simply, just like a cup of coffee and what he least wants is an intellectual hard labor at his desktop.

So as I proposed, we must provide the end user with an absolutely pleasant, *great fun interface*, but behind this interface there must be a *heavy-duty, high performance machinery* to cope with any eventualities *of information retrieval* challenges. This mechanism should allow even the most inexperienced user to search the Web with professional results.

With the advent of "point and click" interfaces, the user-friendly interface exists, on this basis we just have to find further refinements. But as for the background retrieval mechanism, we have to tackle a real tall order. As we rejected the systematic use of universal classification systems, we are obliged to search in another direction. We had also doubts about the capacity of *average* search engines, but *meta search engines* seem to offer a more appropriate solution.

Contrary to what we said about search engines, meta search engines cannot be regarded as databases – they do not even review or classify Web sites or documents, nor accept URL-s. Instead, they simply transmit queries to a great number of search engines at the same time. Given that the doubts still exist concerning the performance of the totality of search engines, the proposed use of meta search engines is far from covering the entire Web, but in the immediate future their application seems to bring considerable improvements and saves us big investments. In fact, however hard and gruesome is to state, we have to realize that *for the time being there is no system capable of reviewing and processing the entire Web*!

As I said, the so called World Wide Web meta search engines (e.g. *Harvester, MetaCrawler, SavvySearch*) dispatch queries simultaneously to the a great number of search engines so that they spare the user the

embarrassment to learn a great many retrieval systems. In fact, the meta search engine replaces effectively what big database providers call *common command language*, enabling the user to search with the same command language in hundreds of databases.

(Some other experts also recommend the use of meta search engines[2] but for me its use is only a "symptomatic treatment" and in another context. In the long term, it will be necessary to synthesize these meta search engines with some leading library classification systems, but this will be an immense operation demanding a huge investment of funds, expertise and manpower.)

So this is the first step to implement a fully customized information management system. The next step is to find mechanisms to provide the user with tailored knowledge.

We need an intermediary mechanism connecting the browser with the meta search engine. The function of this mechanism is twofold:

- translation of the user's rudimentary query into a professionally formulated search strategy,
- provide the user with an interactive interface in order to insure a feedback loop level refinement (or even basic alteration) of the search strategy.

At the core of the system there operates one or a number of thesauri. The thesaurus is the ideal system not only for computerized indexing and retrieval, but for translating the user's inexact query expressions into professional terminology, furthermore, offers an array of tools not only of refining (widening or narrowing) the scope of the search, but also refining its direction towards related and more relevant fields.

So the application of the thesaurus at the core of the background retrieval mechanism makes us available the following tools:

1. *"Transformer"* (not-preferred terms) . It converts rudimentarily chosen, inexact terms and awkwardly formulated expressions into the specific terminology of the retrieval system.
2. *"Telephoto lens"* (narrower terms). It converts too wide search results into adequately narrow searches.
3. *"Wide-angle lens"* (broader terms). It converts too narrow search results into adequately wide searches.
4. *"Steering wheel"* (related terms). It reorients the search direction onto a more relevant route.

It is the task of interface designers to offer the user an easy and enjoyable operation of the above four tools.

An open question still remains concerning the number of thesauri involved at the core of the retrieval mechanism. That is a decision to make by the provider, it depends on the size and purposes of the information services. In this regard it is important to emphasize that even a very small-size provider, but very competent in one specific discipline, is enabled by this pattern to run their independent, reliable, and increasingly appreciated services, and become indispensable on the information market.

Beside the use of terms, the user should be allowed to type in *authors*, *named persons* (about whom the document is written), *corporate items* as well as *time and space* (location) data. The basic *Boolean operators* (OR, AND, NOT) should also be available, preferably in a way that the three *Venn diagrams* appear and one of them will be selected and clicked.

Whether or not one trusts thesauri, they undoubtedly have had a profound impact on the view of the methodological foundations of information retrieval on the Web. In addition, they can be considered as the ultimate quintessence of all sophisticated classification development efforts since Dewey. One can criticize the thesaurus-based approach and state that it is neither new nor original. One can also say that the present revolutionary times demand revolutionary inventions. But can there be discovered something authentically new after one and a half century's classification R & D that has treated all possible aspects and tested them on far vaster areas than the Web? However unpopular it may seem, I must conclude that the basic rules and laws of information retrieval are simple and eternal, therefore one is not allowed to circumvent them.

For optimizing information retrieval accuracy, thesauri should be completed with one very simple but extremely high-performance complement. As I presented, thesaurus-based search expressions consist of a combination of descriptors with some Boolean and proximity operators, but they should *always* be completed with *data that pinpoint the searched information in time and space* (exact date, location plus corporate source, author affiliation). As I tried to prove in a previous paper [3], this method prevents both information surplus and redundancy.

From the viewpoint of user-friendliness, the pinpointing of space and time does not represent any kind of problem. At the same time, the application of proximity operators as well as the use of OR and NOT operators may become very complicated and too cumbersome for the large public, but fortunately, their use is generally not indispensable. I am convinced that the use of descriptors with a symbolized AND, plus the pinpointing of time and space provide in most cases a very easy and effective access to relevant information on the Web.

[1] Zins, C. and Guttmand, D.: (2000) Structuring Web Bibliographic Resources: An Exemplary Subject Classification Scheme. In *Knowledge Organization*. 27. (3) pp. 143-159.

[2] E.g. Thompson, D.M., Egyhazy, Cs. I., and Plunkett, T.K.: (2000). Intelligent Web Search Agents. In *Encyclopedia of Library and Information Science*. Volume 67. Supplement 30. New York – Basel, Marcel Dekker, Inc. pp. 261-262.

[3] Sebestyen, Gyorgy: (1995) Le synchronisme paradigmatico-semantique dans l'analyse documentaire de l'Age electronique. Communication prononcee au 13e Colloque international de Bibliologie de Paris. In *Nouvelles technologies, modèles sociaux et sciences de l'écrit*, Carre des Sciences, 23-26 octobre 1995., Paris, Delagrave, pp. 320-326.