Maria Teresa Biagetti (Sapienza University of Rome)

Pertinence perspective and OPAC enhancement

Abstract
The starting-point of the paper is the debate recently developed in LIS literature about OPAC enhancement and the necessity to design OPACs based on search engines features. Supposed improving tools as relevance ranking and relevance feedback devices are examined. Possible OPAC development lines, based on theoretical examination of relevance and pertinence concepts, according to Saracevic view, and following semantics perspectives, are presented. Finally, enhancement of OPACs starting from their inner characteristics is proposed, and a plan to improve semantic search functions while maintaining existing indexing methodologies, that is document conceptual analysis, is outlined.

1: Introduction
Over the past few years, many scholars in library science have focused their attention on the importance of adapting traditional knowledge organization systems (KOS) to the chances offered by powerful tools as Web search engines. The necessity to design a new OPAC generation, enriched with functions based on search engines features, was clearly identified (Yu & Young 2004; Byrum 2005). Appraised Web search engines features are: easy search, quick show, clear presentation of results and chance of discovering existing kinds of resources.

Effectiveness of bibliographic research by way of OPACs seems really diminished, compared to that achieved by Web search engines and on-line bookshops. It is worth mentioning the OCLC inquiries to analyse users’ bibliographic research behaviour (OCLC 2003; 2005; Weston 2007), which explained that search engines are more frequently used than OPACs. Recently, the Library of Congress has brought into question the suitability for spending enormous budget for keeping complex catalogographic tools, since research engines can retrieve documents using keywords (Marcum 2005).

Among the strategies adopted to conform OPACs to new needs, and to attempt to compete with Web search engines, we can mention the realization of portals, which enable organized access to electronic resources, also through connection with commercial data bases. Further procedures of improvement are: personalization in presentation of services, clustering of users based on their bibliographic research behaviour; the latter uses also data mining strategies (Benoit 2002) applied to information science (Chen & Liu 2004), in particular bibliomining, in order to improve digital library services discovering patterns of use (Nicholson 2006).

A report commissioned by the Library of Congress (Calhoun 2006) has put in evidence that search engines are preferred by “information seekers” and, consequently, suggested neglecting traditional semantic indexing and replacing it with automatic classification. The report recommended integration with other discovery tools and, in particular, the use of keywords organized by relevance ranking, a tool utilized for display of search engines results.

However, it is necessary to distinguish between “quick information seekers”, who can benefit from accessing any information, and “scholars” – in particular, human sciences scholars – who have scientific research needs, for whose satisfaction it is
necessary to implement traditional indexing by using conceptual categories and controlled vocabulary (Mann 2006).

This paper offers a short survey of innovative functions of new generation OPACs and integrated portals and, by approaching the problem of OPAC enhancement from a theoretical point of view, intends to examine the relevance concept taking into account its different aspects. In particular, it suggests to consider pertinence theory as a fundamental paradigm in knowledge organization (KO). A further purpose is to contribute to an examination of OPAC development starting from their inner characteristics, and to propose improving OPAC semantic functions by means of traditional indexing methods.

2: OPAC enhancement

Among the devices proposed to improve OPAC features and make them more similar to search engines, we can examine the use of relevance ranking, that is the presentation of query results by means of a rank, a strategy used by Web search engines, and of relevance feedback.

Relevance ranking is based on a number of factors, such as term frequency and inverse document frequency, as well as the position where a term occurs in the document, the proximity of terms to each other, but also such factors as supposed popularity and interest (Breeding 2007).

Basic relevance ranking is constituted by statistical measures computing the product of term frequency (TF) and inverse document frequency (IDF): TF concerns term frequency within documents, in full-text or in the catalographic records; IDF concerns term frequency in the data base, and if the term occurs in only few documents, its weight is increased. Inverse document frequency is a weight used in order to balance the effect of term frequency. Higher or lower relevance rank can be obtained respectively, if terms in query will be retrieved all together in a field, or each term in a different field.

Examples:

- NCSU (North Carolina State University) Libraries <www.lib.ncsu.edu> have realized a portal providing access to databases, reference tools as encyclopaedias, handbooks, directories, biographies, but also special collections, as manuscripts, rare books, and the university archives. It allows to browse starting from broad disciplinary classes, and searching in NCSU libraries, or in a Triangle research libraries (NCSU, UNC, Duke, and NCCU) or, finally, expanding the search through the access to WorldCat. It is an example of integrated portal, based on the Endeca ProFind™ platform, which uses Endeca Information Access Platform (API) Guided Navigation software, and the MDEX engine, that provides advanced query capabilities to retrieve structured and nonstructured contents. Query results are displayed using relevance ranking as default, and also the date of print is a weight. However, a rank for date alone is also possible. The system permits to refine search results using broad subjects and genres (sources, electronic books, ecc.).
- Libris <libris.kb.se>, the union catalogue of the Swedish academic and research libraries, presents search results using relevance ranking, but no possibility of selecting the title field, or the subject headings field, or anywhere in record, is offered as, on the contrary, it happens in NCSU. In full records users can frequently read tables of contents. Search results are ranked by relevance on the base of words
Queens Library (<www.queenslibrary.org>) is the second largest public library in the U.S., considering the size of collections. AquaBrowser Library platform (<aqua.queenslibrary.org>) permits term search and results are organized by relevance. Besides, the system provides the “word cloud” discovery mechanism, that is a topic map in which, starting from the search term, related concepts, translated terms and spelling variations are disposed. The user can refine search results by using disciplinary classes, series, languages etc. The system permits navigation among terms in titles, in subject strings and in series titles: each term the user will choose in the “word cloud” will be added to the starting term, enriching the knowledge space. Term search may be increased by the “Indexed Content Enrichment” function, which permits searching tables of contents, summaries and fiction profiles for fiction works (<www.aquabrowser.com>).

The adoption of a device as relevance ranking in OPACs should require to provide the explanation of followed principles and criteria, although these criteria are generally not declared.

The Text Retrieval Conference (TREC) (<trec.nist.gov>), assuming the inheritance of C. W. Cleverdon’s work (Cranfield Projects, College of Aeronautics, 1966), from 1992 is aiming to improve retrieval methodologies and new effectiveness evaluation techniques, employing recall/precision measures. In last years (co-sponsored by NIST, DARPA and ARDA), it has expanded its scope, including interactive retrieval, video retrieval and Web retrieval (Harman & Voorhess 2006). For each conference, a set of documents and questions is provided, in order to permit participants to test their own retrieval systems and to prepare a list of the retrieved ranked documents. Then, NIST evaluates the results. The TREC definition of relevance:

“TREC uses the following working definition of relevance: [...] Only binary judgments (“relevant” or “not relevant”) are made, and a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document).” <trec.nist.gov/data/reljudge_eng.html>

2008 TREC Conference has dealt with information seeking behaviour in the blogosphere, information needs of legal community, and exploration of retrieval over a large set of queries and a large collection of documents (The Million Query (1MQ) Track).

Furthermore, TREC 2008 has focused interest on relevance feedback algorithms evaluation and has proposed a common methodology for evaluating and comparing these algorithms (Buckley & Robertson 2008). Implicit relevance feedback techniques work by monitoring users behaviors during their searches, without relevance judgment by the users. Explicit relevance feedback in IR systems is a technique of query-expansion, based on measures of equivalence to a given document retrieved and indicated by the user as relevant. Algorithms of feedback can use similarity in title words, similarity in subject headings, etc. Therefore, the system gives opportunity to retrieve similar documents to that showed as very relevant by the researcher, but similarity is established by the system on the basis of terms in titles or in subject strings.

Turning now to consideration about relevance ranking, the main thing is that it is based on term frequency and it does not satisfy scholars research needs, whose interest is in conceptual context. As Thomas Mann has pointed out, scholars

“also wish to avoid having to sort through huge lists or displays – from any source – in which relevant
materials are buried within inadequately-sorted mountains of chaff having the “right” keywords in the wrong conceptual contexts.” (Mann 2006, 8)

3: Relevance/pertinence
That means there is some evidence that scholars’ real problem is to retrieve documents in which concepts are regarded in accordance with the particular significance (sense), aspects or contexts they want.

Let us consider thoroughly the concept of relevance. relevance has been considered for thirty years as the theoretical foundation of information science (Saracević 1975). Actually, two different concepts of relevance were analysed. The first one, “subject knowledge view of relevance”, concerns the relation between a query and knowledge and literature about a topic, considered as public knowledge. The second, “pertinence view of relevance”, is concerned with the connections that appear between personal knowledge of a researcher and the public knowledge. Recently Saracević (2006) has confirmed his view of relevance, and the distinction is now between topical relevance (or subject relevance), and user relevance. The latter is subdivided in situational relevance (or utility), the relation between informative objects and a particular situation or a task, and cognitive relevance, or pertinence, the relation between informative objects and the status of knowledge of a user at a particular time.

It is worth mentioning also that Douglas J. Foskett (1972) suggested to consider relevance as a kind of public knowledge, a model of thought universally accepted. In that case, the reference was to scientific paradigms theory by Thomas Kuhn (1962).

According to Saracević, pertinence concerns personal knowledge and the knowledge models in single researchers’ mind. It concerns the nature, structure and “sedimentation” of user knowledge. Pertinence is a relative peculiarity, not a persistent characteristic attributed to a document once and for all. Pertinence relation is established every time within the framework of the single search.

An IR system can answer a query, but can just guess what is the information need of a user. Relevance: queries; pertinence: information need.

The information need concerns lack of knowledge. A system returns items matching relevance, that result pertinent from the point of view of a particular researcher, but a system can also return items matching relevance that will not be considered pertinent from the point of view of that researcher; on the contrary, there could be items considered pertinent by the researcher, not returned as relevant by the system (Fugmann 1993).

4: Semantic perspective
The assumption of pertinence as a basis for a feasible enhancement of OPAC semantic functions links up with the proposal of reconsidering future developments in information science in accordance with the semantics point of view (Hjørland 2007). According to Hjørland, it is of great significance to acknowledge that IR tools, as thesauri, ontologies, and bibliometric maps too, show a semantic nature. Among semantic research devices, we can consider also the potentialities offered by co-citation and bibliographic coupling measures, which permit to quickly realize semantic linking among documents:
“I hold that the citing relation is in itself a kind of semantic relation. In support of this claim, I distinguish between “ontological” and social semantic relations and argue that citing relations belong to the latter.” (Hjørland 2007, 394).

Hjørland assumes a pragmatist view of semantics, following the theories expressed by the American philosopher Hilary Putnam, focused in particular on the functional-pragmatic nature of meaning. The center of interest is the consideration of meanings and semantic relations relativity.

“Semantic relations relate to a given task or situation and not all users of a given set of semantic relations will share the same view of which terms are equivalent. On the other hand, it is clear that if we base a semantic theory on an individualistic/idiosyncratic view of concepts and semantics, it is not possible to design systems for more than one user or situation-an absurd conclusion.” (Hjørland 2007, 381)

A broader philosophical framework for the comprehension of the semantic approach can be recognized by mentioning the science of signs devised by Charles Sanders Peirce (1931-35; 1938) and the pragmatics point of view, according to Charles Morris’ elaboration (Morris 1938). In Morris’ view pragmatics concerns the study of the relations among signs and their interpreters. As part of semiotics, pragmatics examines sociological and psychological aspects involved in the process of signification.

5: Plausible developments

As an alternative to a wide adoption of devices as, for instance, relevance ranking, or relevance feedback, which permit a too modest improvement of semantic research capabilities, a likely enhancement of OPACs could consist in an enrichment of their semantic searching functions, which could occur from inside, expanding potentialities and inherent capability for their growth.

Considering that semantic relations are potentially countless, and that information systems, and OPACs, cannot understand what “information need” a user has every time, and which document can serve the purpose of a specific research, adopting the pertinence perspective, it could be a good improvement to let the users know the variety of subjects a document presents, and further, what depth, richness or complexity in subject treatment documents present.

The enhancement can be achieved using traditional semantic indexing strategies. The proposed model can be summarized along these lines:

(1) Conceptual analysis of documents devoted to differentiate two perspectives and, consequently, to put in practice two different indexing strategies every time it is needed, for instance in case of ideologically biased books:
   (a) to identify the object(s) a document is dealing with. As an example: For a document dealing with heresy (arianism, catharism, and so on), the object could be heresy (or heresies)
   (b) to identify the subject(s) a document is dealing with, that is considering the discourse that an author communicates about that topic. As an example: a document could present a critic speech about heresies, from a catholic point of view, for instance, and the subject could be catholic orthodoxy. This could safeguard pertinence.

(2) Thorough strategies devoted to the conceptual analysis of documents (books, articles), which could permit to put in evidence arguments dealt at different
levels, that is main (or principal) arguments and minor (or secondary) arguments in the document. As an example: a book dealing with absolutism in 17th century can present also minor arguments, as history of finance, or social conditions.

(3) Different headings for representing main and minor subjects in the research system. The use of different marks, or different size of letters, or different colours, will permit immediately to differentiate them, understanding the level of depth of the subject.

The plan could permit to maintain traditional indexing methodology, that is conceptual document analysis, and to provide an articulated system that consist of objects and main and minor subjects, relating to different degree of breadth.

Using visual interface, as that, for instance, offered by the AquaBrowser Library platform, which permit to graphically distinguish terms, with a differentiation of different degree of subjects treatment, the user could search within strings only terms referred to objects or to main subjects.

References
Foskett D.J., 1972, A note on the concept of relevance, Information storage and retrieval, 8, p. 77-78.
Fugmann R., 1993, Subject analysis and indexing: theoretical foundation and practical advice, Indeks, Frankfurt M.
OCLC, 2005, Perceptions of libraries and information resources: a report to the OCLC membership, Online Computer Library Center, Dublin (Ohio).

*Web documents have been accessed 28 September 2009.*