The Dynamic Nature of Searching and Browsing on Web-OPACS: The CATHIE Experience

Abstract: The paradigm shift from the old system centered view to a user centered approach involves new tools needed for accessing library resources under the condition that the user's needs are taken into account. An end-user, who has only a little knowledge of classification systems or thesauri, understands little of the mode of the representation of contents and the use of authority lists. In addition, he will have difficulty in formulating his question in a precise manner. He needs to know better what the library proposes in order to define of what use it would be for him. Many studies have been carried out on the use of controlled vocabularies (classification, authority lists, thesauri) as searching devices. It is surprising to find that relatively little attention has been given to the role of these tools in filtering and browsing processes. We have developed a prototype named CATHIE (CATalog Hypertextuel Interactif et Enrichi) that supports such filtering and interactive reformulation features.

1. Introduction

The paradigm shift from the old system centered view to a user centered approach involves new tools that are needed for accessing library resources under the condition that the user's needs are taken into account. Classical models of information retrieval interact only minimally to describe the dynamic nature of the interaction. Recently different models have attempted to describe further the ways in which information retrieval interaction is dynamic. These models include Saracevic’s (1997) stratified model, Belkin’s (1996) episodic model, Bates’(1989) berypicking model and Ingwersen’s (1996) global model of polyrepresentation. Bates (1989) argues that classical IR models assume that a user’s information need is constant. Techniques like relevance feedback enable the user to build up a syntactic representation of a single static query. In contrast, she observes, many users begin with one aspect of a broad topic and continually develop new conceptions of their underlying information need. The users find results of interest at many of the steps taken in information exploration. We have reason to assume that seeking information is seldom completed with only one single set of queries and one or more retrieved documents. More often, the initially retrieved set serves as feedback for further query formulations. Knowledge organisation can play an important role in these processes.

Actually, the central problem of present online systems is improving precision. Most people using IRS try to find information using one or two word queries. As a consequence, the retrieval from actual keyword IRS will often generate a very large volume of hits, frequently too large to be of any practical use (Ihadjadene, 1999; Silverstein,1998). One of the most radical effects of keyword searching in online catalogs is to increase recall. Moreover, studies on tactics used show that, when there are too many hits, browsing the search results is the most popular tactic used, followed by the addition of terms to the first search statement (Bruza, 1997), (Ihadjadene, 1999).

In this paper, we seek to improve this situation (information overload) through a user centred approach that provides users with tools for narrowing search results. One approach is to offer a broad range of retrieval functions and mechanisms for controlling the retrieved results. The basis for this study was the question about whether current classification systems are suitable for filtering search results.
2. Searching and Browsing Strategies in CATHIE

According to Marchionini (1995), a major goal of research and design in HCI and IR is to determine what strategies are useful and how they can be supported by new systems. One approach is to use online searching strategies identified in studies of professional online searchers to create systems that optimize those activities. Examples of strategies are: "building blocks approach", "facet strategy" and "item instantiation strategy". In CATHIE, we had developed the latter two.

The item instantiation strategy is proposed in most operational WWW-OPACS. In a previous study Ihadjadene (1999) examined the searching and browsing behaviour of Enssib-users. The finding was that the use of syntagmatic relationships (subject headings) in browsing Web-Opacs is a popular strategy both for local and remote users (14.05 % vs 16.37%). Users generally browse by using subject headings (70.31 % to enssib vs 43.87 % to Lyon2), and use the author links (19.65 vs 30.96%) to a far lesser extent. In information retrieval systems that allow for feedback processes, the assumption is that objects from the subject file (keywords) are only of importance for reformulation of the queries. Our results reveal the potential for the use of other sources of terms in the reformulation. Our results show that the users, while browsing, continually redefine the information problem, and that in many cases, it is the user's desire for redefinition that initiates and drives the browsing process. In CATHIE, we have limited browsing by author and subject terms (see figure 2).

We have also implemented the facet strategy in CATHIE. This strategy enhances the development and construction of complex Boolean queries. Facet strategy begins with a large subset of the database and successively pares it down with facets specific to the problem. According to Harter (1986), this strategy works well with problems that are vague or broad. Our notion of facet strategy is also similar to the Infoseek model. Infoseek allows users to limit a collection of results by issuing a query over the already existing query. This action is equivalent to issuing a query that is a conjunction of the query that created the first collection and the new filter query. One difference is that CATHIE offers a set of terms for narrowing the search results (see figure 1). These terms were extracted using a statistical analysis algorithm.

3. Reformulation in CATHIE

An end-user, who has only a little knowledge of classification systems or thesauri, understands little of the mode of representation of contents and the use of authority lists. In addition, he will have difficulty in formulating his questions in a precise manner. He needs to know better what the library proposes in order to define of what use it would be to him. One approach that has been traditionally regarded as a tool for increasing precision is the use of interactive reformulation. One way is to offer an analysis of a set of retrieved records. Statistical analysis will quickly show term distribution and may suggest terms the searcher had not considered. This functionality was found to be available in some commercial systems such as ESA (ZOOM), Dialog (RANK), Orbit (GET) and Questel (MEMS).

The ZOOM command counts the number of occurrences of unique terms within a specific field from an established answer set. The terms are ordered from the most frequent to the least frequent and then placed in alphabetical order if the count is the same. This is also available in some experimental online catalogues such as OASIS's "summarize" function which lists all subject headings in retrieved sets ranked by decreasing frequency of occurrence (Buckland,1992). Efthimiadis (1993) conducted a series of evaluations to see which sorting function worked best for interactive reformulation. One search produced results. CATHIE was used to analyze the 100 most relevant records, to extract the subject headings from each record retrieved and to present the headings in order of frequency. CATHIE displayed only the seven most relevant subjects. If the number of documents retrieved is larger than 30,
CATHIE downloads the complete number of subject headings (main headings with the subdivision) found in MARC 6XX fields. If not, it downloads only the main headings.

After a request on "droit et information", CATHIE displays the seven most pertinent subject headings to the users. With this list as a background, the user can reformulate his query. Also, the general categories (classes and sub-classes of DDC) are displayed (figure1). This enables users to break down large sets of retrieved documents. This feature for filtering concept terms is important because recent studies showed that only a small percentage of users take advantage of the query refinement mechanisms (Spink, 1999; Dennis, 1998).

4. Classification and Filtering

Many online catalogues offer search by classification number. It is usually found that people make very little use of this option. Typically fewer than five per cent of searches use this approach. The main reason was that classification code numbers are not meaningful to most users. In 1983, Svenonius highlighted a number of areas in which searching could be enhanced using classification. It could be used to improve the recall and precision of searches, to save users time in inputting search terms, to contextualize the user's search, and to allow for the search to be broadened or narrowed. Many studies have been carried out on the use of vocabulary control (classification, authority lists, thesauri) as a searching and browsing device. It is surprising to find that relatively little attention has been given to the role of classification in filtering processes.

We know, that the tasks of browsing and reading the items retrieved from a screen is not a trivial one. Exploration can be time consuming as soon as the number of potentially relevant records becomes large. The lists of retrieved records from IRS systems lack organization beyond ranking. Studies by Wiberley (1990) and Spink (1999) speak about using online catalogs and search engines to show that most users look only at the first few results in the list. The ranking of search results is not sufficient. In order to address this problem, many alternative output displays have been proposed by the HCI searchers. Visualization interfaces such VIBE, Scatter/Gather or Cat_at One allow the user to browse simultaneously through a large hierarchical category and a set of documents. These systems make use of 3D technologies for this display (Korfhage, 1997).

In library science, some researchers have recommended the assignment of broad category codes to enable systems to summarize large numbers of subdivided subject headings by broad subjects. Markey (1996) demonstrated how classification schemes can be used to manage highly posted searches. Larson (1996) promotes the use of LCC to cluster bibliographic records in OPACs when sets retrieved from subject searches are large. They provide users with an overview of the range of items retrieved, organizing them by subject. Using classification to filter or partition results was also proposed by Chan (1995). A number of others researchers suggested that there be automatic links between subject headings and class marks to allow users to browse subject headings in a systematic, rather than an alphabetic, manner. The exact mechanism for providing this in online catalogs was explored in two works Micco (1991) and Hildreth (1993).

The CATHIE system that we have developed allows users to dynamically organize a retrieved set according to multiple factors such of publication date, language and the topic of a document. When the title keyword search on “access” is run against the CATHIE database, 49 items were retrieved. The results are presented to show the different DDC categories these records fall into, allowing a user to see the various contexts in which the terms have been used.

NortherLight uses classification data-mining by grouping search results into “folders” that have been created by librarians. The NortherLight system shares with our approach the goals of making the filtering of query search results easier. It differs in its focus on documents rather than on concept terms.
The weakness of this hierarchical filtering is the restriction that one specific term may appear in only one category. The monodisciplinary classification character is responsible for this weakness. The intellectual province of general bibliographic classification is the whole universe of knowledge and this domain has habitually been analyzed into classes and subclasses on the basis of the academic disciplines. As Beghtol (1999) demonstrates, the present general classifications of knowledge do not adequately support multidisciplinarity topics. The principal reason is that the classifications serve primarily as tools for shelf arrangement. However, assigning multiple class numbers to items and exploiting the Dewey structure in a facet form can improve the CATHIE filtering mechanisms (Tinker, 1999).

4. Conclusion

We have designed an experimental online catalog that can improve browsing and filtering compared with using a classical Boolean (or statistical) system alone. It assists users in organizing sets of documents. The features of CATHIE can be implemented easily in any operational system. CATHIE brings together a number of current strategies in a single system. The CATHIE prototype is still, in its current state, experimental and more technical improvement can be added:

-We can for instance, adapt and test CATHIE in different specialized fields of knowledge (medicine, patents, law, etc.). As we observed during its development, CATHIE depends basically on indexing systems linked to bibliographic IR systems. We would simultaneously make use of specialized thesauri (MeSH, INSPEC, etc.) and other classification systems to help descriptors and document filtering.

-We can develop a WWW version of CATHIE. We are looking forward to developing a search engine that will facilitate filtering and interactive bibliographic information filtering. The W-Cathie Prototype will be developed in Java and will make use of the Z39.50 standard to access distant catalogs. This solution would not be feasible without a re-examination of the compatibility of indexing languages.

A next step would be testing the effectiveness of this approach in quasi-controlled retrieval tests with online catalogs users.

References


Votre question :

1. Droit
2. Information

Classification des résultats par thèmes :

(21 rép.) Droit
(7 rép.) Sciences et Techniques

Résultats de votre recherche : 32 réponses

4. Vigue, Basque (BP) - Droit à l'information, 1980.
7. BP - Droit de l'Information, 1924.
8. Jura (BP) - Droit et propriété, protection. Licence, 1924.

Figure 1: filtering in CATHIE

Votre question :

1. Auteur
2. Titre
3. Editeur
4. Langue
5. Sujet

Résultats de votre recherche : 23 réponses

6. Sujet, Systèmes d'information - Evaluation

Figure 2: Browsing in CATHIE