Information Coverage – Incrementally satisfying a searcher's information need

H.A. Proper

Ordina Institute
Groningenweg 6
2803 PV Gouda
The Netherlands
E.Proper@acm.org

Th.P. van der Weide

University of Nijmegen
Toernooiveld 1
6525 ED Nijmegen
The Netherlands
TvdW@cs.kun.nl

Published as:

H.A. Proper and Th.P. van der Weide. Information coverage - Incrementally satisfying a searcher's information need. In C.~Stephanides, editor, *Universal Acces in HCI: Towards an Information Society for All*, pages 719--722, New Orleans, Louisiana, USA, August 2001. Lawrence Erlbaum Associates, Mahwah, New Jersey, USA. ISBN 0805836098

1. Introduction

The Internet has become the virtual reality of mankind – a world that we shape without many of the imperfections of reality. We can jump to literally every place in no time and reach every resource anywhere anytime. In particular this last promise of *information at your fingertips* is under siege. The growing complexity of information space overwhelms the wired consumer and the vast increase in information is outpacing the improvement of retrieval tools.

Traditional information retrieval systems aim to satisfy a searcher's information need by matching some explicit formulation of the searcher's information need to the set of available information carriers (such as documents, web-pages, etc.) and then returning the set of information carriers that best match the formulation of the searcher's information need. From the perspective of the searcher, this traditional approach has two serious drawbacks:

Need formulation: Searchers are presumed to have a very clear understanding of their information need, even though it is not likely that they will be able to articulate their precise information needs in terms of a query language. In the case of the Internet, this becomes even more apparent as the collection of available information is sheer endless.

Need satisfaction: Even when the entire set of returned information carriers is indeed relevant to the searcher's information need, they are still required to manually wade through the result sets in search of the right combination of information carriers to cover their information need.

In traditional information retrieval systems, searchers are not provided with an advice on an effective order to best read (a selection of) the information carriers to cover their information need. The word 'effectively' may refer to time, financial costs, low overhead, etc.

Other drawbacks have to do with the quality of document characterization and limitations of matching algorithms. On the next few pages we report on ongoing research. This research extends on work as reported in [Proper and Bruza, 1999; Weide et al, 1998]. The current state of the research is that we have (formally) defined the information coverage problem area in the style of [Proper and Bruza, 1999]. As a next step, we are developing strategies and algorithms to actually develop an (idealised) information portal.

2. Information coverage paradigm

In our research, we take the perspective that searchers are not just looking for a set of relevant information carriers, but really expect an information retrieval system to help them in covering their information need in an effective way. It is our belief that an information retrieval system should really play the role of an information portal as illustrated by the information coverage paradigm as shown in Figure 1.

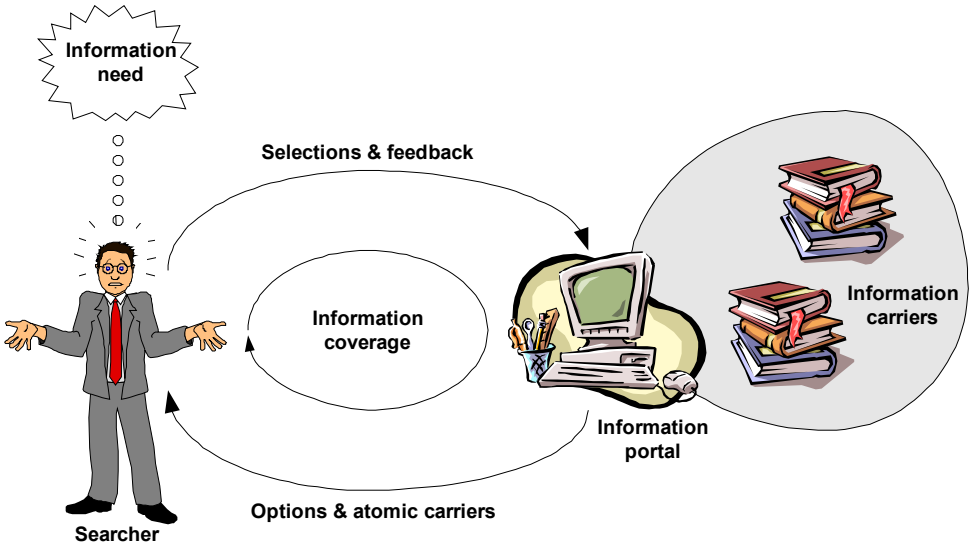

Figure 1: The information coverage paradigm

On the left hand side there we find a searcher who is in need of information. Facing the searcher, on the right hand side, we find a collection of available information carriers. This may be a limited collection of information carriers that are available in some library, but could also be a sheer endless set of carriers, such as all information available via the Internet.

In the middle we find an *information portal*. The information portal aims to satisfy the searcher's information need. It tries to do so not only by communicating with the searcher in terms of queries and result sets. An information portal should really communicate with searchers on two levels, rather than one.

This first level is the *clarification level*, focusing on the clarification and discovery of the searcher's actual information needs. At this level of communication, the *need formulation* issue as discussed above plays a crucial role. The communication between the searcher and the information portal should ideally be seen as an interactive dialogue in which the information portal tries to clarify the precise information need while incrementally satisfying this need. The concept of *query by navigation* as discussed in, e.g. [Bruza and Weide, 1990] and [Bruza and Weide, 1992], is an example of such an approach.

The second level is the *information level*, which is concerned with the actual information aimed to fulfil the searcher's information need. At this communication level, the *need satisfaction* issue is a major challenge. Communication with respect to the actual information the searcher is looking for involves the information portal providing the searcher with information carriers, and some form of (relevance) feedback from the searcher to the information portal. This feedback may pertain to the informational content, as well as aspects such as the medium used, length, genre, etc. Ideally, the feedback from the searcher to the information portal should be as

complete as possible with regards to the perceived relevance to the searcher's information needs. The more feedback the information portal receives from the searcher with regards to the relevance of information carriers, the better the information portal will be able to tune the set of selected information carriers to the actual information need.

The combination of both levels of communication, with the intention of covering a searcher's information need, is what we refer to as *information coverage*. In our research we aim to develop a theory for this information coverage process, where we take the functionality of an *idealised* information portal as a starting point. The ensuing challenge is then to identify an as-large-as-possible subset of functionality of the idealised information portal that can actually be realised.

3. A modest view on searchers

Since we are ultimately aiming for an *automated* information portal, we have, even at the level of the idealised information portal, made some pragmatic assumptions with regard to cognitive and psychological aspects of searchers. These assumptions focus on the goals with which searchers turn to the system as well as the state of mind of the searchers themselves.

For an (idealised) information portal, it is imperative to understand the goal and expectations with which a searcher turns to the system. We presume there to be two different levels for defining the goal of searcher.

The information need with which a searcher turns to the information portal is the searcher's *cognitive goal*. In addition to the cognitive goal, a searcher is likely to have some *operational goal* as well. This goal relates to the tasks the searcher has to perform. Tasks, which have led the searcher to turn to the system in the hope to gain new knowledge relevant to the task.

While, as discussed above, it may not be an easy task for searchers to express their cognitive goal, it will be less hard for them to express their operational goal. Especially when this can be done using some pre-defined terminology in the context of the searcher's task description. A context that, for example, may be provided by a business process model or a workflow model. Operational goals can have different forms, e.g. task profiles of employees, student goals when attending courses, etc.

Finally, searchers are assumed to be 'objects' that may be in different states; the searcher states. In each of these states the searcher will be in some mood. A mood that is likely to influence the searcher's ability to take up new information from information carriers. Furthermore, in each state the searcher is viewed as having some (active) body of knowledge at their disposal; their knowledge in that particular state. In general, the intention of searchers, when turning to an information portal to satisfy their needs will be to achieve a particular change in their mood and/or to acquire additional knowledge.

We realise that the above assumptions leave out numerous intricate aspects of human cognition and psychology. However, we believe that our limitations would still yield an information portal that would be more attuned to the needs of searchers than existing approaches offer. Using the above modest view on searchers, our model of an idealised information portal may already seem to be too ambitious to automate.

4. A theory for demand and supply of information

The theory for information coverage uses the metaphor of infons [Barwise, 1989; Proper and Bruza, 1999] as a semantic base. Infons are used as a representation form of elementary information particles.

In general, the intention of searchers, when turning to an information portal to satisfy their needs, will be to achieve a particular change in their mood and/or to acquire additional

knowledge. Information carriers are the vehicles by which the searcher may satisfy these intentions. Experiencing (e.g. reading, listening or viewing) an information carrier will bring a searcher in a different state. A number of assumptions describe the nature of the state change function. For example, it is assumed that experiencing an information carrier does not lead to a loss of pre-existing knowledge. What information the searcher actually absorbs, depends on the mood of that searcher.

Depending on how information from information carriers is absorbed, a number of typical searcher profiles can be defined. An example is a maximal absorber, being a searcher who will absorb all information available from an information carrier on the first time they experience this carrier.

In our research, we initially restrict ourselves to the need for knowledge, which is effectuated as a need for infons. This need is communicated to the system in a searcher-independent style, for example as a query. The actual demand of the searcher thus depends on the current state of that searcher. When matching this demand for information with the information supply from the information carriers available, a classification can be made. For example, a distinction can be made between the new information supplied by some information carrier, the overhead information that would be supplied by some information carrier, etc.

In small search tasks, a stable need for information may be assumed. However, in the general case, there may be a demand drift. For example, when reading a document on a certain topic, this may raise a further interest in some aspect of the original information need. Furthermore, the structure of compound information carriers is to be taken into account as well.

5. Cost models

When a collection of alternative information carriers is available to a searcher, some mechanism is needed to determine the relative relevance of these carriers to the searcher's needs. One factor in determining this relevance is the topical relevance of the information carrier with respect to the searcher's information need. This is, however, not the only factor involved. Other factors, such as:

- the reliability of the information provided by an information carrier,
- the amount of money that may have to be paid to obtain the carrier,
- the cognitive load the searcher must endure in reading the information carrier,
- the amount of time needed to gain access to the information carrier,

may have to be considered as well. To be able to rank the available information carriers, while taking such factors into account, a price/performance ratio is introduced. The better the price/performance ratio of an information carrier, the more preferred the carrier is presumed to be. The actual price/performance ratio associated with a set of information carriers is based on some underlying cost model. As 'electronic commerce' has increasingly gained attention, the need for such models becomes more pressing [Wondergem et al, 1998].

A complicating factor is the fact that the price/performance ratio cannot be calculated for one particular information carrier in isolation. Some of the factors involved in computing the price/performance ratio do not behave monotonous with respect to sequences of information carriers. An example of such a factor is the purchase price of an information carrier. When obtaining more information carriers from the same supplier, subsequent information carriers may, for instance, be obtained at a lower price then when they were purchased in isolation. To illustrate the effects of non-monotonic behaviour, consider the following example:

Suppose chapters A-1 and A-2 are from book A and chapter B-1 is from book B, while A-1 and B-1 are similar in content. Let us also presume book A is priced at € 40.- while book B is priced at € 30.- and that the searcher does not yet own either other the books.

If the information portal would have to chose from either A-1 and B-1 to relinquish the searcher's information need, the portal is likely to opt for B-1 as this requires the purchase of book B, which is cheaper then book A.

However, if A-2 is also needed to cover the searcher's information need, then first purchasing book B might not be such a good idea after al, as it is still required to also purchase book A.

One might argue that in the age of E-Commerce, the above example is antiquated as one may choose to use micro-payments to purchase specific chapters of even smaller pieces of relevant information. However, even in the case of micro-payments, the purchase price may still behave non-monotonous. It is not unlikely that content providers will try and increase the loyalty of their clientele by offering price reduction based on a customer's volume of consumption. In that case, it may be wiser to purchase chapters A-1 and A-2 using a single content provider, rather than purchasing B-1 and A-2 from different content providers.

6. Further research

It should be the intentions of any information portal to find those information carriers that best correspond to the information requirements of a particular searcher, within the specified cost restrictions. This goes beyond the traditional retrieval task, as it also requires the system to monitor searcher behaviour, such as some approximation of the searcher's state (covering both mood and knowledge).

The incremental model has been introduced [Weide et al, 1998] to model partial satisfaction of the information need as an effect of supplying an information carrier to the searcher. Extending this model with a cost function provides the opportunity to solve simple versions of the information discovery problem. A typical application is to find a minimal cost set of information carriers that completely cover some information request.

The incremental model can be introduced within the context of the model presented in the previous sections. Currently we are looking for strategies to map the requirements of the 'idealised information portal' as discussed above on to the concrete operators as introduced in the incremental model.

Bibliography

[Barwise, 1989] Barwise, J. (1989). The Situation in Logic, CSLI Lecture Notes.

[Bruza and Weide, 1990] Bruza, P. and Weide, T. v.d. (1990). Two Level Hypermedia - An Improved Architecture for Hypertext. In Tjoa, A. and Wagner, R., editors, *Proceedings of the Data Base and Expert System Applications Conference (DEXA 90)*, pages 76-83, Vienna, Austria. Springer-Verlag.

[Bruza and Weide, 1992] Bruza, P. and Weide, T. v.d. (1992). Stratified Hypermedia Structures for Information Disclosure. *The Computer Journal*, 35(3):208-220.

[Proper and Bruza, 1999] Proper, H.A. and Bruza, P.D (1999)., What is Information Discovery About?, *Journal of the American Society for Information Science*, 50(9):737-750.

[Weide et al., 1998] Weide, T. v.d., Huibers, T., and Bommel, P. v. (1998). The Incremental Searcher Satisfaction Model for Information Retrieval, *The Computer Journal*, 41(5):311-318.

[Wondergem et al, 1998] Wondergem, B.C.M, Bommel. P. v., Huibers, T.W.C. and Weide, T. v.d. (1998). Opportunities for Electronic Commerce in Information Discovery. In F. Griffel, T. Tu, and W. Lamersdorf, editors, *Proceedings of the International IFIP/GI Working*

Conference on Trends in Distributed Systems for Electronic Commerce, TREC 98, pages 126-136, Hamburg, Germany.