Guided Information Exploration

Jesus Favela
favela@cicese.mx
CICESE Research Center
Km. 107 Carr. Tijuana-Ensenada, 22860, Ensenada, B.C., México

Abstract

A model for retrieving non-solicited information is presented. The model is based on the guided exploration of information spaces. It works by monitoring user actions and formulating a hypothesis of her information needs. This hypothesis is used to locate documents similar to those which the information seeker has found to be interesting. The manipulation of these documents by the user is then used to update the hypothesis of information needs and retrieve a new cluster of documents that the system estimates as being relevant to the user's information needs. We present a general model for guided information exploration, discuss the functional requirements of the main components of the model, and propose an architecture for implementing it.

The Process of Guided Information Exploration

Information plays an important role in problem solving. It allows us to improve the solution to a problem or simply makes it feasible to arrive at a solution. In the midst of this process, however, we often ignore that we need information, we are unable to articulate our information needs in the form of a query or we are simply unable to locate this information. In these scenarios we are often assisted by collegues, advisors or other professionals who provide us with un-solicited information after they have formed at least a vague idea of our information needs from a brief conversation with us or by just looking at the type of material that we read or the actions we take to solve a problem.

In this paper we present a model for retrieving information relevant to the problem at hand based on a similar process. In this case, the computer acts as an assistant that "looks over our shoulders" at what we do and the type of information we focus on and discard. Based on this information, the system creates hypotheses of our information needs and "suggests" to us information relevant to these needs.

We have named this process Guided Information Exploraton (GIE). The dictionary gives the following definitions of exploration and guidance:

Explore. To traverse over a region for the purpose of discovery.

Guide. To accompany (a sightseer) to show him points of interests and to explain their meaning and significance.

Two aspects from the previous definitions are significant to us. First, exploration (the process of exploring) is a purposeful action that denotes a desire to discover something, although it is not precisely known what this discovery will be or where will it eventually take you. Second, guidance implies assistance in directing the search process and in making sense of the information being accessed. For the purpose of supporting this activity, we propose the following definition of Guided Information Exploration:

> "An iterative strategy for accessing information in digital form by approximating user information needs in the form of hypotheses, retrieving documents relevant to these needs and updating these hypotheses based on feedback provided by the user's manipulation of the information that was previously suggested to him, until the user is satisfied with both the formulation of his needs and the information retrieved."

Figure 1 shows the cognitive tasks that an information seeker is involved with during the process of accessing information. First, the user needs to realize that he needs information, and that this need can be satisfied. Second, he approximates his information needs in the form of a question or query. Third, the user interprets the results of the retrieval presented by the system. After these results are interpreted and evaluated, if the user is not satisfied with this information, he can either modify the query to fine-tune the search, or he can use this information to reformulate his information requirements. Declarative query languages such as SQL have simplified the formulation of queries. The later problem, that of modifying information needs in the light of new information, is what information exploration is aimed at. In the extreme case, when the formulation and subsequent changes to

the information needs are not done by the user directly, but by the system (based on implicit information obtained by the monitoring of the user actions) is what we call guided information exploration. Since the user is not explicitly asking for information we say that the system is retrieving unsolicited information relevant to the problem at hand.

Guided information exploration is an appropriate paradigm for accessing information in circumstances that have the following characteristics:

- The user does not know that he needs information or if he does, he does not know where to find it or how to ask for it. Ironically, it is precisely when we dont have the knowledge required to access relevant information when we are in greater need of this information.
- The user is trying to solve a problem[1]. Accessing information is not his primary activity. The value of information can be determined by its effects on the decisions that are made to solve the problem.
- The amount of resources that can be invested in accessing information is limited. If the cost of accessing information is high, the user will make a decision with the information he already has. At issue is whether or not the value of additional information is worth the cost of accessing it.

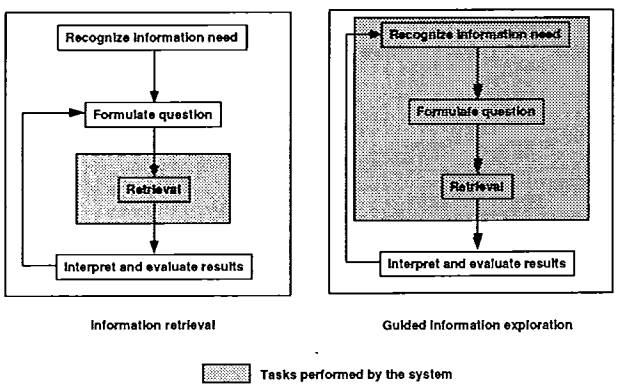

Figure 1. Cognitive tasks involved in accessing information

Figure 2 illustrates the general model of the guided information exploration process. The model consists of a module that monitors the user's manipulation of documents and creates hypotheses of his information needs; a second module that uses these hypotheses to retrieve relevant documents; and a presentation system that allows the user to browse the documents being suggested by the system. We now discuss the characteristics of each of these modules.

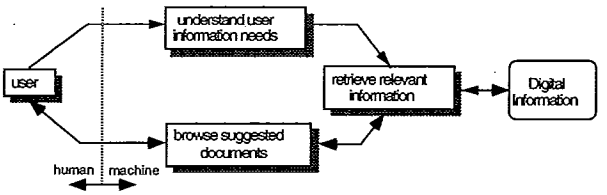

Figure 2 A model of the guided information exploration process

Creating hypothesis of user information needs

The formulation of hypothesis of user information needs is based on the monitoring of the user manipulation of electronic documents. The argument for using document manipulation as an information source to formulate a query is our hypothesis that sufficiently clear semantics can be identified for these operations in terms of the interests that the user has on the information that is being presented to him. For example, clicking a hot region of an image, and with it opening a detailed view of that image, reflects an interest on that specific, and in general similar, information. Closing a document, on the other hand, indicates that the user has lost interest on the information contained in this document.

In recent years, social approaches to cognition have emerged that oppose traditional interpretations of social action. Among these approaches is the field of Ethnomethodology, an approach to sociological studies advanced by Garfinkel (Heritage 1984). In contrast with traditional approaches to the study of social phenomena which take social facts as objective truths, in ethnomethodology the existence and creation of these acts becomes the subject of study. As explained by Lucy Suchman (Suchman 1987), the objective of ethnomethodology is "how it is that the mutual intelligibility and objectivity of the social world is achieved" (p. 58). Garfinkel has argued that social interaction is possible because we are able to interpret and create expectations of the actions of others, thus creating a "shared understanding" that allows us to communicate with each other. We can, for instance, cross the street without being threatened by a coming car because we expect the car to stop when the red light is on. If it were not for these interpretations of the actions of

[1]Here we use the word "problem," to refer to what Simon calls "wicked" or ill-structured problems (Simon 1969). In contrast with "scientific" problems, the solution to an ill-structured problem is not true or false, but rather, good or bad. The difference is important because the value of information is measured in terms of how it improves a given solution.

others, we would be overwhelmed by the number of things that we would have to consider before making a simple decision. Suchman has applied these ideas to the problem of human-machine communication where the interpretation of user actions generates a response from the system. An interface should be "interpreted" by its user.

The language of a computer application is defined by its interface. As a user manipulates this interface, a kind of conversation is established between the user and the machine. Direct manipulation interfaces make this language more explicit by creating metaphors of real-world scenarios that are familiar to the user, and which include operations that can be directly invoked by the user to create objects, change their attributes and establish relationships among them. In direct manipulation interfaces, the semantics of user actions are made explicit and can be the subject of interpretation.

In the case of a system used for information browsing, a direct manipulation interface is responsible for controlling the way in which the information is displayed, and provides the means by which the user can select and operate on the information on which (s)he is interested. The actions supported by the documents and the presentation primitives used to display information can give a quantitative measure of the interest that the information seeker has on those documents and form the basis for the creation of a model of user information needs.

An interaction language defines a set of subjects, actions, and rules that are used to form correct statements for human-machine interpretation. The subjects of manipulation are documents (layouts of presentation primitives) and the views of information objects (objects that contain information and support operations to display this information). Actions for the direct manipulation of documents are of two types: window manipulations, such as opening and iconizing a window, and hypermedia operations, such as following a link or marking a node. Media specific operations can be applied to the views of information objects, examples include playing an audio segment, stepping to the next frame of a video clip or scrolling text. The following are two examples of interaction statements in textual form:

move window "Alternative decision" to (0,70)
follow link "module 3" in current document

To create an hypothesis of user information needs we have to identify those actions that when applied to a document reflect the users navigation through information nodes, and from which his information needs can be at least partly inferred. Once these actions have been identified, rules for the construction of correct statements are defined for each interaction class.

The module responsible for understanding user needs and formulating hypothesis of these needs should satisfy the following requirements:

Modularity.
Ideally, relevant information must be accessed while in the midst of problem solving. This suggests that the system that supports GIE must be ubiquitous and integrated with the application that the user is using to document or support his decision making process. If the problem is that of scheduling a set of tasks, for instance, the user should be provided with relevant information while using a scheduling package, rather than having to move from one system to the other, or wait until the problem has been partially solved to ask for relevant information.

A GIE module must be designed in such a way that it can be "attached" to different applications to support the retrieval of relevant information in the domain for which the application is being used. In the case of a system that supports product development, this information could include design principles, handbooks, and previous projects. A complete solution thus includes an application that supports problem solving, information relevant to that type of problem, and a generic mechanism that performs hypothesis formation, retrieval, and presentation of relevant information.

Use of implicit information.
In order to afford information exploration, a user must be relieved from the cognitive load of explicitly giving feedback about the relevance of the information that is presented to him. Therefore, implicit methods must be used to extract this information. As we suggested above, the interpretation of user actions is one such method. Explicit feedback could be used if an appropriate interface is provided that does not distract the user from interpreting the information that is being suggested.

The retrieval of relevant information

This part of the exploration process involves identifying the documents that more closely satisfy the hypothesis of information needs that are produced from the monitoring of user actions. To accomplish this, the system should be able to estimate the relevance of each of the documents in the database to a given hypothesis. This can be done directly by comparing each document to the query or this comparison can be accelerated by, for instance, first clustering all the elements in the database and then retrieving those documents that belong to the same category as that in which the query is placed. We now discuss two of the most important requirements of the retrieval mechanism.

Fuzzy match.

Databases normally use boolean or exact match to determine whether a given document is relevant to a query. In this case, the set of retrieved documents is said to be crisp since a document is either relevant or irrelevant to a given query. Specifically, in querying, information needs have to be formulated precisely and with no ambiguities. Imagine for instance a business application, where a corporation might want to notify all its employees whose salary is more than $100K of a new tax policy that affects them. In this case the boundary of the set that differentiates these people from other members of the organization is very clear, and a database can be used to obtain a list of members of this group.

In information exploration, on the other hand, information needs are not well defined, and might in fact evolve as the user gains better understanding of the problem at hand. In this case it makes sense to talk about the degree of relevance that a document has with respect to a given information need. If the mechanism used to retrieve relevant documents can estimate the degree to which a given document belongs to the set of relevant documents, this information can be used to rank these documents in order of relevance.

Adaptable to continuous input.

An important aspect of the information exploration process is its iterative and dynamic nature. In GIE the documents that are suggested as being relevant should constantly be updated to reflect the changing information needs of the user and the convergence of the approximation of his information needs by the system. This requires an information retrieval engine that can work with a continuous flow of information requests and that can quickly update the list of relevant documents.

The hypotheses of information needs used in GIE are refined incrementally. They suffer slight modifications each time, especially at the end of the process when the hypothesis converges to the real needs. Therefore, an efficient retrieval engine does not have to make a complete search every time the hypothesis is modified, but should incrementally update the list of relevant documents.

The presentation of relevant documents

Once relevant documents are retrieved they are presented to the user for their evaluation. If guided exploration is supported as a module in an application used for problem solving, the application will be responsible for the display of the documents that are found to be relevant. Still, the module responsible for presenting these documents should be subject to the following functional requirements:

Non-intrusive.

If the user is in the midst of solving a problem he might not want to be interrupted by the display of the documents that the system finds to be relevant. This is particularly true for documents that have low relevance. In an information accessing system that uses approximations of information needs for retrieval, it can be expected that a large percentage of documents retrieved will have medium to low relevance. This is not necessarily bad since it is precisely in this region where relevant but non-obvious information can often be found (Hiltz & Turoff 1985). However, the interface should allow the user to quickly browse the documents being presented so that (s)he can discriminate the relevant ones without being overwhelmed by the amount of information.

In information retrieval systems there are two measures commonly used to determine the effectiveness of a retrieval method or systems: *precision* and *recall*. Recall measures the percentage of the relevant documents in the database that are retrieved by the system, which can be interpreted as the probability that a relevant document will be retrieved. Precision is the percentage of the documents retrieved that are relevant, which gives an indication of the quality of the information retrieved or the probability that a document that is retrieved will be relevant:

$$Recall = \frac{No.of documents retrieved that are relevant}{No.of relevant documents}$$

$$Precision = \frac{No.of documents retrieved that are relevant}{No.of retrieved documents}$$

In a well known study, often cited as an indication of the limitations of current information retrieval technology, Blair and Maron (Blair & Mason 1985), evaluate a state of the art information retrieval system using a large database of legal documents and obtain an average recall value of 20% and precision of 79%, which leads them to conclude that the technology, which works well for small databases, does not scale up. In this study, however, a successful retrieval session is defined by the users as that which obtains at least 80% recall. Other applications, such as the filtering of electronic mail messages, might require much lower values of recall. In GIE we are interested in accessing information that might have a positive impact on problem solving and in this regard some relevant information might be better than none at all. It might be wrong to expect, therefore, that a GIE system will retrieve most of the information relevant to a given topic.

As illustrated in Figure 3, for a particular system or retrieval technology, precision and recall have an inverse relation, that is, one can increase the number of relevant documents retrieved (more recall), but this will normally involve getting more documents that are not relevant as well (less precision). A good overall measure of the effectiveness of a retrieval system is the precision-recall product. It is expected that

better information retrieval technologies will drive the precision-recall curve away from the origin, indicating better precision-recall products. Given that GIE uses vague and ambiguous representations of information needs, we are already resigned to having low precision and recall values, even if we use the best information retrieval methods available. Precision should, however, improve as the exploration proceeds and the requirements become more specific.

The key to accepting a relatively large number of irrelevant documents is the time associated with browsing and assessing the relevance of these documents. The interface used to present these documents can help in this regard if it allows users to easily distinguish irrelevant document. We can use the metaphor of reading a trade magazine to explain this argument. When a reader selects a magazine to read he makes a statement about the general topic he is interested in. Even though he does not know the details of the information he will find, and given that easily more than half the content will be irrelevant (advertisement, for instance, accounts for more than half of the content of a computer magazine), he still decides to browse the magazine because the cost of reviewing and discarding a page of irrelevant information is very low.

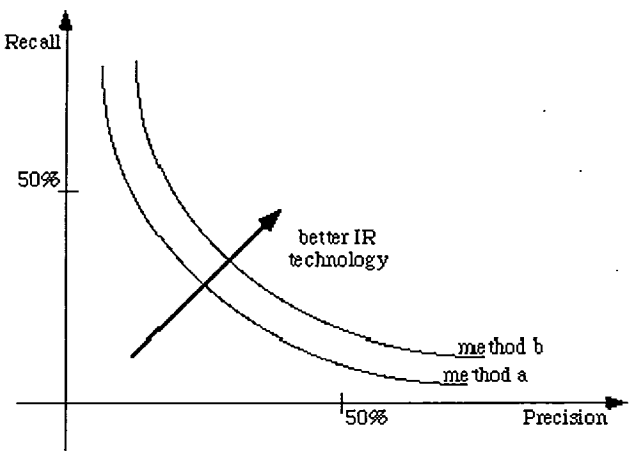

Figure 3 Precision and recall in information retrieval systems have an inverse relation, increasing one normally decreases the value of the other.

Similarly, the presentation of documents in GIE must allow the user of the system to quickly assess the relevance of a document and discard it if necessary with an effort equivalent to that of flipping a page in a magazine.

Display of multimedia data.

Documents should be displayed using the media that is most natural for the information being presented. In doing this, it helps information seekers make sense of this information and accelerates the capture of information.

Supporting the display of multimedia objects also involves supporting operations relevant to the different media being used. One should, for example, be able to select a segment of an audio clip and play only this segment, or expand an image to look at some of its details. Somewhat related to this issue is the ability to select segments of relevant information and annotate and store them in a personal notebook for future reference.

Implementing Guided Information Exploration

In this section we present an architecture for implementing Guided Information Exploration (GIE). This represents just one alternative implementation that satisfies the functional requirements discussed above.

The GIE process is illustrated in Figure 4. The figure also shows the implementation of GIE as a service in an application that supports problem solving. The first step in the implementation consists of identifying the user actions on documents of the application that give some indication about the users interests in the information contained in that document. To estimate the user interests in a document given the evidence provided by the manipulation of that document a backpropagation neural network (Rumelhart *et al.* 1986) is used. The input to the network is a vector $\{e\}$ that represents the actions performed on the document j, the output to the network is a single value $r_j \epsilon [0,1]$ which indicates the relative importance that the document has to the user given the evidence $\{e\}$. This network is trained with a set of input and output patterns obtained from the interaction of actual users of the system, where they are asked to estimate the relevance of a given document after a period of interaction with it. This training set is used to associate user commands with document relevance, a mapping that will be generalized by the network and which will allow the system to estimate the relative importance that a document has to a user from the way in which the document is manipulated.

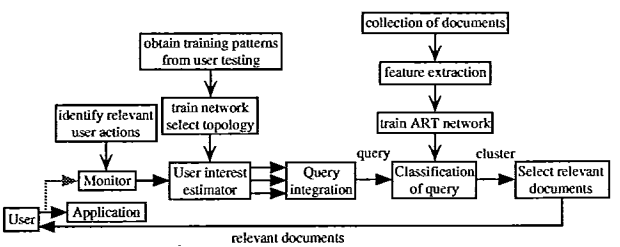

Figure 4. Steps in the methodology to implement GIE (vertical flow), and to use GIE (horizontal flow)

The retrieval process can be accelerated with the use

of a Fuzzy-ART (Adaptive Resonance Theory) neural network (Carpenter *et al.* 1991) which classifies the documents in the database prior to the presentation of a query. The process then consists of finding the cluster to which the query belongs and estimating the similarity between the query and the documents in the database that belong to this cluster. Those documents with similarity higher than a given threshold are then suggested to the user as being relevant to his/her information needs. The process continues with the monitoring of the users reaction to the documents being presented, in order to update the hypothesis of information needs.

The key steps in the implementation of a given application consists on identifyning user actions on documents that given an indication of the user's interests in the information contained in these documents and training the backpropagation network to obtain the relative importance of a document given a series of manipulations performed on it.

Conclusions

We have presented a model for retrieving non-solicited information based on the guided information of exploration spaces. We have identified the main modules of such a model, its functional requierements, and an architecture for its implementation.

The implementation was tested with a small (32 cases) database of user interface design cases. The signature of each document was composed of the functional requirements and design decisions of a specific user interfaces. The system retrieves design cases similar to the one being designed thus suggesting design alternatives and solutions to past problems similar to the ones being addressed.

Guided Information Exploration could allow members of an organization to take advantage of a pool of loosly indexed projects that form an important part of the organization's memory.

References

Blair, D. and Maron, M. 1985. An Evaluation of Retrieval Effectiveness for a Full-text Document Retrieval System. *Comm. of the ACM* 28(3): 289-299.

Carpenter, G.; Grossberg, S.; and Rosen, D. 1991. Fuzzy-ART: Fast Stable Patterns by an Adaptive Resonance System. *Neural Networks,* 4: 759-771.

Hiltz, S.R., and Turoff, M. 1985. Structuring Computer-Mediated Communication Systems to Avoid Information Overload. *Comm of the ACM* 28(7): 680-689.

Heritage, J. 1984. *Garfinkle and Ethnomethodology,* Cambridge: Polity Press.

Rumelhart, D.; Hinton, G.; and Williams, R. 1986 Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing,* Vol. 1, Rumelhart and McCelland (eds.) Cambridge, Mass.: MIT Press.

Simon, H. 1969. *The Sciences of the Artificial* Cambridge, Mass.: MIT Press.

Suchman, L., *Plans and Situated Actions: The Problem of Human-Machine Communication,* Cambridge University Press, 1987.