Expert/Consultation System for a Retrieval Data-Base with Semantic Network of Concepts

Peretz Shoval
Graduate School of Business
University of Pittsburgh, PA 15260

Abstract

This paper describes a development and implementation of an expert/consultation system for a retrieval data-base, that interfaces between the user and a retrieval system. The system's objective is to perform the information consultant's job in assisting a user to select the right vocabulary terms for his query. It is particularly useful for a novice user of a controlled-vocabulary, index-based retrieval system, who is not familiar with the vocabulary and the system Thesaurus. The user will enter his terms/key-words, that represent his information need, and the system will apply search procedures on its knowledge-base, and will find relevant concepts to be used as query-terms. The system is interactive; it can explain to the user why/how a concept was discovered/suggested, and it can backtrack and try to find alternatives in case the user rejects a suggested concept. Two versions of the system were developed, utilizing two search and interaction strategies. Experiments will be conducted with the two alternatives in order to find out user preference and to compare performance. Performance will also be compard with an alternative "conventional" approach, which is an On-Line-Thesarus - developed as part of this study.

The Problem

Performance of a retrieval system depends on many factors. An important one is using the right set of key-words to represent the information need/query. It is both the problem of the indexes and the searcher. This study takes the searcher's point of view. In a controlled-vocabulary, index-based retrieval system, key-words represent the contents, and access to the data is done via these key-words. The searcher (and also the indexer, in turn) confronts the central problem of which key-words to use to formulate his query.

A user of a retrieval data-base can usually consult two sources for assistance in shaping his query:

a. The Data-Base Thesaurus; which usually has several indices, like: Main (alphabetic) Index, with cross-reference relationships, Hierarchical Index, and KWOC, or Rotated Index. Although Thesauri usually provide instructions of how one should use them. There is no doubt that a 'random' user can not make an effective use of it (to be convinced, just have a look at how complicated are these instructions).

b. Information Consultant/Analyst; who is supposed to bridge the gap between the user needs and the system. The user expresses his problems (in writing or orally) and the consultant conducts the search of the Thesaurus to find the right terms.

   In order for him to perform effectively, an expert has to be available to the user, be familiar with the area of application and its concepts, be experienced with the Thesaurus and its controlled vocabulary,and be familiar with the vocabulary that might be used by various/potential users of the data-base-users who do not necessarily use vocabulary concepts. In other words; the consultant has to understand the meaning of terms used by various users, and then find the right vocabulary terms to express the query. If a user happens to be assisted by a consultant with these skills, he is likely to get good advice and hence also good results from the data-base (in terms of recall and precision). The problem is that experts with these qualifications are not always available.

©1981 ACM 0-89791-052-4/81/0500-0145 $00.75

One can find an analogy between the job of the information consultant and the medical doctor (an analogy which is supported in the literature). A doctor is doing (in first step) a medical dianoigis, which means: define what the problem is--according to symptons and other facts. Similarly, the information analyst has to define what the information need is--'normalized' into vocabulary terms. In both cases, we are at the 'conceptual' stage of the 'problem-solving' process of an ill-structured problem.

Principles of the Expert System

A computerized expert/consultation system ought to perform the job of a human-expert. A human-expert has the knowledge in the subject area (whether he remembers or has access to knowledge) and working methods/procedures that he applies on the knowledge in order to find the solution for a given problem. In other words, the expert takes the facts (given in the problem), and applies his work methods (which is his expertise) on the world of knowledge--to find the solution. Our expert system utilizes this general scheme with the two components: knowledge and procedures:

Knowledge is represented as a semantic network; where nodes are terms in the subject area, and links are the various types of relationships between them. The source for this semantic-net is the data-base Thesaurus: its terms, whether in controlled vocabulary or not, are the nodes. The cross-reference relationships are the basic links. We usually distinguish between BT/NT links, that represent hierarchical relationships, USE/USED-FOR, that represent synonymous relationships, and RT, general relatedness. To these 'basic' relationships we add more types of relators, to represent additional knowledge that an expert may have:

a. 'Generator' links, which combine source/generic words and terms to multiword concepts. For example: the concept "information-system" may be generated by the generic terms "information" and "system," and the concept "management information system" may be generated by the concepts "management" and "information systems." These combinations create a hierarchy. They enable access from generic words to vocabulary terms which have a narrow meaning. A user who is not familiar with the concept "management information system" may opt to use, for example, "information systems for management." The "generate" relationships will help to find the right concept. The function of this relation is similar to the service provided when using the KWOC/Rotated Index, in the Thesaurus.

b. "Model" links: every concept in the network is viewed as a model, which means: links to its various meanings and components. For example, the concept "Business," in addition to its "regular" cross-reference relationships and "generate" links may link to components like "organizational structure" or "functional areas." These concepts describe the concept of "Business." Each of them, in turn, may be described by linking to its own components. In other words, the 'Model' links extend the knowledge about concepts, beyond the Theasurus definitions.

The second component of the expert system is the procedures/search algorithms. We apply a search process in the network, where the objective is to match/intersect concepts. This is one of the principles of a human-expert's work: he takes user terms and tries to intersect them by finding concepts with a narrower meaning. Hence, if a match is found, the new concept is identified instead of its originators. Another principle is expansion: if a match is not found, we try to expand the concept along its various links to other concepts, to see if any of its constituents can be matched against any other concept. At the same time, we look for synonyms, so that non-vocabulary terms will not be suggested. Not all 'matched' terms are considered potential concepts; we apply some rules to limit the consideration.

Control of the Expert System

Two alternative search and interactive strategies were developed; one called "Interactive Strategy" and the other "Best-First Strategy."

The control of the 'Interactive-Strategy' version of the system is as follows:

- The system accepts user terms, and immediately expands them along the semantic-net links.
- The new/expanded concepts are matched against all other concepts. If an intersection is found it is first verified that the match also involves additional entry terms, and if it does - it is suggested to the user, who is asked to judge whether or not the suggested concept is in right direction. Hence, with this strategy the user is the 'evaluation function.'
- If the user rejects a suggested concept, as being in the 'wrong direction' - search in that direction stops. If he accepts it - search continues on from that node, since the direction is 'promising', so that this node is further expanded. Hence, we use a 'best first' search strategy. As a concept is accepted its 'parents' are marked (all the way back to its originators) and are excluded from the current list of suggested-concepts. Hence, only the 'front-line' concepts, which capture the meaning of their 'parent-concepts' are considered.
- The process continues as long as intersections are found; and then the search proceeds by expanding the concepts which were not matched - trying to match their constituents - and so on.
- Finally we end up with a list of 'suggested-concepts,' for each of which the user is given the following options:
  - He may accept (again) the concept. In this case the system will simply continue to suggest the next-best concept.

- He may ask why/how a concept was suggested. The system will printout its Parent-Tree - tracking back to the originators.

- He may reject a suggested concept, in which case the system will backtrack and try to find alternative concepts. (This backtracking process is recursive, and may continue all the way back, as long as the user continued to reject alternatives.)

The order in which concepts are suggested to the user depends on 2 factors:

a) How many user/entry terms were involved in its creation, which is an indicator for its importance or acceptance. Concepts with more originators are considered 'better'.

b) How may additional entry terms the concept carries, as compared to the entry terms that were already considered by the previously accepted concepts. This means: consider the best complement for the already accepted concepts. This factor makes sure that the user is presented with a cluster of concepts which as a whole represent the meaning of his query.

If we combine both factors we find that at the beginning a set of concepts which represents the whole user-query, is suggested, and later the system may continue to suggest additional concepts, if it has more.

The second version of the system, 'Best-First Strategy,' performs the first stages of the search process without interacting with the user: It assumes that the intersections found during the search are accepted. During the second stage the user is presented with the same options as before, and if the system's assumption were wrong - the user can 'fire' the backtracking procedures, which try to find alternatives.

Experiments and Further Research

Both versions of the system will be compared in a set of experiments, to find out user preferences, differences in performance (comparing the sets of suggested terms) and difference in terms of 'interaction-load' (comparing how 'fast', in terms of interaction, the system reaches the results).

In addition, these systems will be compared with a 'conventional' approach to consulting. For this purpose an alternative system, namely an On-Line Thesarus was developed. This alternative system accepts a user term and presents to him cross-reference relationships of the term to other terms in the knowledge base.

The system described in this paper was implemented, using a data base to store the semantic network. It is only a first step. In further developments the system should be connected to an actual retrieval data-base, so the user will be able to get feedback and correct the query accordingly. With a larger implementation of the knowledge base it will be also possible to compare its performance with a human consultant. Other steps in the development of this system should include query-formulation, and entry of user query in natural language, instead of keywords.

The printout presented on next two pages show an example of the system in work, using the 'Best-First - Strategy'. The key words entered by the user in that example represents the following query: "What methods are there for user/management evaluation of satisfaction with retrieval information systems?"

Acknowledgement

The system presented in this paper was developed as part of my Ph.D. Dissertation. I wish to express many thanks to my advisor, Dr. Harry Pople Jr., for the advice and support provided by him during this study.

Selected Bibliography

Beck, C. McKehnie, T. and Peters, P.E.: Political Science Thesaurus II, American Political Science Association and University of Pittsburgh, 1979.

Debons, A.: "The Information Counselor," Journal of Information Counselor, Vol. 1(1), Jan 80.

Lancaster, F.W. and Fayen, E.G.: Information Retrieval on Line, Meville Pub., L.A., 1973.

Minker, J.: "Information Storage and Retrieval - A Survey and Functional Description," ACM/SIPIR, Fall 1977.

Pople, H.E., Jr.: "Heuristic Methods for Imposing Structure on Ill-Structured Problems: The Structuring of Medical Diagnostics," in: Artificial Ingelligence in Medicine, American Association for the Advancement of Science, Aug 1980.

Pople, H.E., Jr. and Decision Systems Lab: "The Formation of Composite Hypotheses in Diagnostic Problem Solving - An Exercise in Synthetic Reasoning," Fifth IJCAI, 1977.

Quillion, R.M.: "Semantic Memory," in Information Processing, Editor: M. Minsky, MIT Press, 1968.

Salton, G.: The SMART, Retrieval System: Experiments in Automatic Document Processing, Prentice Hall, N.J. 1971.

Smith, L.C.: "Artificial Intelligence in Information Retrieval Systems," Information Processing and Management, Vol. 12, 1976.

Woods, W.V.: "What is a Link: Foundations for semantic networks," in: Bobrow and Collins (eds.): Representation and Understanding, Academic Press, 1975.

```
                        ***WELCOME TO THE SYSTEM***


 SELECT AN ALTERNATIVE, AS FOLLOWS:
            INTERACTIVE-STRATEGY    -ENTER "I"
            BEST-FIRST-STRATEGY     -ENTER "B"
>B


**STEP-1. "BEST-FIRST STRATEGY"

PLEASE ENTER TERMS, ONE BY ONE. AT END ENTER "DONE"

>METHOD
>EVALUATION
>USER
>MANAGEMENT
>SATISFACTION
>RETRIEVE
>INFORMATION
>SYSTEM
>DONE
         8 TERMS ENTERED

**STEP-2. "BEST-FIRST STRATEGY"

  I WILL NOW PRESENT TO YOU MY SUGGESTED TERMS (AND,IN PARENTHESES,HOW
 MANY ENTRY TERMS CONTRIBUTED TO IT).

  FOR EACH SUGGESTED CONCEPT YOU HAVE THE OPTIONS TO:
*ACCEPT IT (AND I WILL CONTINUE TO NEXT CONCEPT)-             SAY "YES"
*REJECT IT (AND I WILL BACKTRACK,TRYING TO FIND ALTERNATIVES)- SAY "NO"
*SEE WHY/HOW IT WAS SUGGESTED (AND I WILL PRINTOUT
                               ITS SOURCES/"PARENT-TREE")- SAY "PP"
*ASK FOR CONCEPT EXPLANATION/DEFINITION-                      SAY "EXP"
*PRINT THIS MESSAGE AGAIN-                                    SAY "HLP"


-COMPUTER SYSTEM MANAGEMENT              (3)>NO


-MIS                                     (3)>PP
 --MANAGEMENT INFORMATION SYSTEM
   ---INFORMATION SYSTEM
      ----SYSTEM

      ----INFORMATION
   ---MANAGEMENT


-MIS                                     (3)>YES


-BENEFIT COST ANALYSIS                   (3)>PP
 --PERFORMANCE EVALUATION
   ---EVALUATION
 --ECONOMIC MODEL
   ---MODEL
      ----METHOD
      ----MANAGEMENT SCIENCE
          -----MANAGEMENT


-BENEFIT COST ANALYSIS                   (3)>YES
```

```
-ISRS                                          (3)>PP
 --INFORMATION RETRIEVAL
   ----INFORMATION
   ----RETRIEVE
 --COMPUTER BASED INFORMATION SYSTEM
   ----INFORMATION SYSTEM
      -----SYSTEM
      -----INFORMATION


-ISRS                                          (3)>YES


-GOAL SATISFACTION                             (2)>NO


-USER                                          (1)>YES


-SATISFACTION                                  (1)>YES


*AT THIS POINT I HAVE CONSIDERED ALL THE TERMS YOU ORIGINLLY ENTERED.
 IN ADDITION, DO YOU WANT TO EXAMINE SOME MORE RELATED CONCEPTS ?
 ENTER "YES" OR "NO">YES


-AUTOMATED AUDITING SYSTEM                     (3)>NO


-COMPUTER BASED MANAGEMENT CONTROL SYSTEM(3)>NO


-MEASUREMENT                                   (2)>PP
 --METHOD
 --PERFORMANCE EVALUATION
   ----EVALUATION


-MEASUREMENT                                   (2)>YES


-SYSTEM PERFORMANCE                            (2)>YES


-LIBRARY RESEARCH                              (2)>NO


-SIMULATION                                    (2)>NO


**NOW, TO SUMMARIZE, HERE IS THE FINAL LIST OF
  SUGGESTED & ACCEPTED CONCEPTS:

 -MIS                                           (3)

 -ISRS                                          (3)

 -BENEFIT COST ANALYSIS                         (3)

 -MEASUREMENT                                   (2)

 -SYSTEM PERFORMANCE                            (2)

 -USER                                          (1)

 -SATISFACTION                                  (1)

   *** THANK    YOU ***
```