Chapter

FAQ-Centered Organizational Memory

Shih-Hung Wu, Min-Yuh Day, Tzong-Han Tsai, Wen-Lian Hsu
Institute of Information Science
Academia Sinica
Nankang, Taipei, Taiwan. R.O.C.

Abstract: The value of a piece of information in an organization is related to its retrieval (or requested) frequency. Therefore, collecting the answers to the frequently asked questions (FAQs) and constructing a good retrieval mechanism is a useful way to maintain organizational memory (OM). Since natural language is the easiest way for people to communicate, we have designed a natural language dialogue system for sharing the valuable knowledge of an organization. The system receives a natural language query from the user and matches it with a FAQ. Either an appropriate answer will be returned according to the user profile or the system will ask-back another question to the user so that a more detailed query can be formed. This dialogue will continue until the user is satisfied or a detailed answer is obtained. In this paper, we apply natural language processing techniques to build a computer system that can help achieve the goal of OM.

Key words: Ontology, Question-answering, Information map.

1. INTRODUCTION

Knowledge collection and documentation within a large organization is a critical issue. It is equally important to make the stored knowledge easily accessible within the organization. We shall address the following questions: What kind of knowledge needs to be preserved? How do we store the knowledge? How do we utilize the stored knowledge? To deal with the above issues, we have designed a knowledge representation system and a natural language dialogue system. We believe that the value of a piece of information in an organization is related to its retrieval (or requested) frequency. Therefore, collecting the answers to the frequently asked

questions (FAQ) and constructing a good retrieval mechanism is a useful way to maintain OM.

We use the term "FAQs" in a broader sense. Take the knowledge related to a department of customer service as an example. Our FAQs shall include all kinds of questions that are being asked to that department plus the additional ones that are potentially important. We treat the answers to these FAQs as valuable pieces of knowledge to the organization, which need to be maintained. These FAQs are indexed in our knowledge representation map and their corresponding answers are stored in databases in a distributed fashion. An answer can be a document, a diagram, a program, a database, a video or an audio recording. To make the system more user friendly, there should be as few restrictions as possible on the format of the OM representation. However, unlike pure text data, different media of data storages do not have a uniform management environment and there is no easy way to retrieve the desired answer. Hence, the key to the retrieval problem lies in an effective indexing mechanism. In this paper, we use a "natural language question" as an index for each knowledge piece rather than a detailed form-based description as seen in most metadata approaches.

The knowledge representation map in our system contains the ontology of the domain knowledge plus necessary linguistic knowledge for matching a user's natural language query with a correct system FAQ. The answers of these FAQs can be maintained by domain experts who do not need to be knowledgeable about the map. The map construction only has to deal with the question portion of these FAQs, which is likely to be more manageable. Our hierarchical knowledge representation scheme allows proper representation of knowledge at different resolutions. For information not represented in natural language, our system shall index them in natural language and enable them to be easily retrieved.

The notion of corporate or organizational memory has been discussed for over a quarter of a century. OM is defined as the means by which knowledge from the past is brought to bear on present activities, thus resulting in higher or lower levels of organizational effectiveness [Stein 1995].

Different users may use different ways to state the same question, the main task of the interface is to map all these different surface forms to one representative question, which can then be used for the retrieval of the FAQ knowledge base [Winiwarter 2000]. There are several related QA systems and dialogue systems, such as the Murax system [Kupiec 1999], which determines from the syntax of a question if the user is asking for a person, a place, or a date. It then tries to find sentences within encyclopaedia articles that contain noun phrases that appear in the question, since these sentences are likely to contain the answer to the question. The idea of the FAQ matching system [Burke et al. 1997] is to match question-style queries

against question-answer pairs. The system may use some standard IR techniques to find the most likely FAQ pairs for the question and then matches the terms in the question against the question portion of the question-answer pairs. There are some systems [Jocobs 1993, McCune 1985, Strzalkow1999] attempting to parse natural language queries in order to extract concepts to match against concepts in the text collection. We use our knowledge map as a basis to parse the queries. Thus, Our system's ability to parse depends more on the system's knowledge than on grammar.

2. ONTOLOGY AND NATURAL LANGUAGE QUERY UNDERSTANDING

We have implemented a dialogue system to demonstrate the idea of FAQ-Centered OM. The system consists of a knowledge representation scheme INFOMAP, a knowledge editor, a natural language query matching mechanism, a solution editor, and a FAQ-based dialogue mechanism. Figure 1 shows the architecture of our QA system.

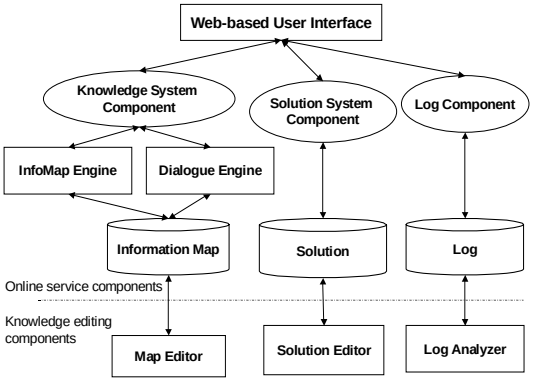

Figure 1. The system architecture

2.1 Our Knowledge Representation – INFOMAP

In this section we introduce our knowledge representation scheme that matches a natural language query to a query concept, where a query concept is extracted from the question portion of a FAQ [Hsu 1999, 2001]. Our knowledge representation scheme, INFOMAP, provides a mechanism to represent and utilize the knowledge. INFOMAP has a hierarchical tree

structure, though this is only a deceivingly simple statement since it does contain “references” that connect nodes on different branches, which makes it into a more general “network”. The root node is usually the name of a domain or a subject such as passport, department store. Following the root node, the first level nodes down are topics that users may be interested in. These topics have sub-categories that list related sub-topics. The knowledge representation has taxonomy similar to the directory structure in most Internet yellow page. However, besides these concept nodes, INFOMAP contains many function nodes, which provide different relationships between the related concept nodes. The addition of appropriate function nodes will produce a much powerful result for the search.

For example, figure 2 shows the related information of libraries (in Academia Sinica) and some FAQs. It can be seen from the figure that the information of a library is divided into three function nodes: category, event and attributes at the first branch under the library agent. The category of library consists of a list of libraries such as chemistry library, earth science library, Chinese literature library, European and American studies library, life science library, economics library, and Information Science library. The attributes of library include admission, research sources, document collection, latest news, web databases, regulation, electronic journals, etc. The event of library can be “borrowing books from the library”. Nodes under the “Category” form a hierarchy, which is the taxonomy of library. Each node under “attribute” forms a new hierarchy (which is not part of the taxonomy of library). The nodes under “Event” are similar. The “FAQ” function node marks an end of a query concept. All the nodes from the root to the “FAQ” form a path, which is also a query concept. The nodes on the path are the key concepts of the FAQ. Under “FAQ”, there is a function node “condition”, which serves as a variable of one of the key concepts. The “condition” can be matched to different examples of a concept, which makes the representation of FAQs more flexible.

Consider the FAQ, “The regulations of library,” under the node of regulation. There is a condition node “library” in the figure. This condition refers to any specific library under the category of library. Thus, the function node “condition” acts like a variable to represent these libraries in related question. The synonym of regulation can be “The policy of using library” or “using library” or a template. The template here means “the regulation of”(optional) + “borrow” or “borrowing” + ”a book” or “books”. This template can be used to match various sentences that have the same query concept. Therefore, question like the ones listed in the test function nodes “How can I borrow a book from the Chemistry Library”, “The regulation of using IIS library”, and “What is the regulation of using Life Science Library” can all be mapped to the same FAQ.

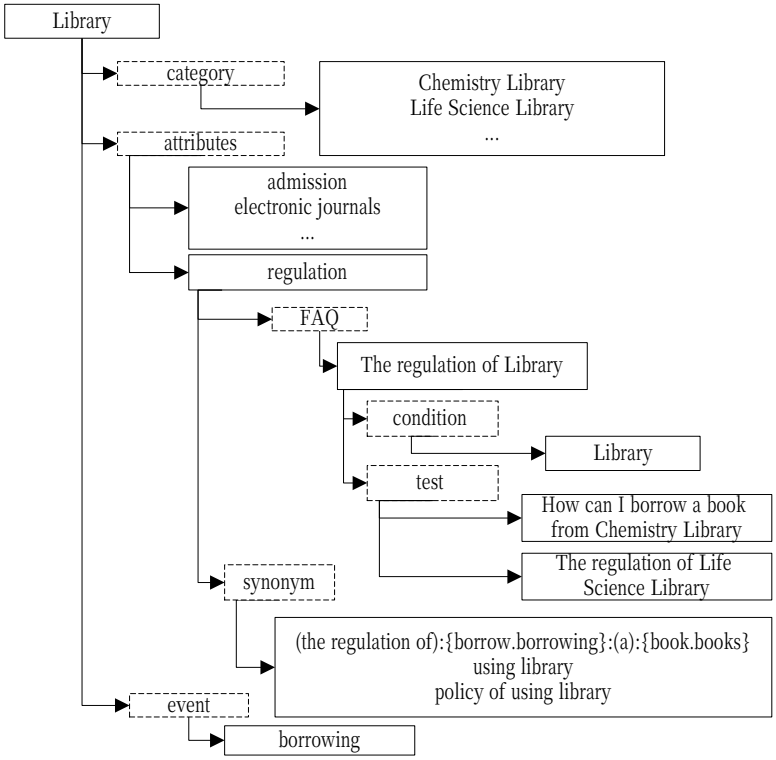

Figure 2. A partial view of the INFOMAP for the concept "library"

2.2 Function Nodes Indicate Relationships

In INFOMAP, we use function nodes to label the relationships between two concept nodes and their hierarchies. The basic function nodes are: category, attribute, example, synonyms, event, FAQ, condition. These function nodes help to represent and identify query concepts.

1. Categories: Various ways of dividing up a concept A. For example, for the concept of "people", we can divide it into young, mid-age and old people according to "age". Another way is to divide it into men and women according to "sex", or rich and poor people according to "wealth" and etc. For each such partition, we shall attach a "cause". Each such division can be regarded as an angle of viewing concept A.
2. Attributes: Properties of concept A. For example, the attribute of a human being can be the organs, the height, the weight, the hobbies and etc. To facilitate linguistic analysis, we also attach the semantic categories of each noun according to various semantic trees.

3. Associated events: Actions that can be associated with concept A. For example, if A is a “car”, then it can be driven, parked, raced, washed, repaired and etc.
4. Synonym: Expressions that are synonymous to concept A.
5. Examples: Instances of A. For example, if A is “hotel”, then its examples can be the actual hotel names.
6. FAQ: A typical question associated with the concept node.
7. FAQ conditions: Items in a query that can be substituted by examples of concepts such as cities, hotels and etc.

The list of function nodes can be expanded for different applications whenever necessary. Each node B underneath a function node of A can be treated as a related concept of A and can be further expanded by describing other relations pertaining to B. However, the relations for B described therein will be “limited under the context of A”. For example, if A is “hotel” and B is the “facility” attribute of A, then underneath the node B we shall list those facilities one can normally find in a hotel, whereas for the “facility” attribute of a specific hotel A1, we shall only list those existing facilities in A1. There are also other function nodes that are used specifically for linguistic feature extraction and parsing, such as articles, modifiers and the semantic categories of these modifiers. Syntactic constraints on A in a phrase or sentence are also considered.

2.3 INFOMAP Serving As Ontology

Ontology represents domain knowledge and serves as a common understanding of the domain. Ontology consists of definitions of concepts, relations and axioms [Guarino 1999, Staab 2001].

Knowledge is stored and used according to a specific ontology for each knowledge management system. The ontology can be very simple such as keyword hierarchies or rather complex such as the WordNet [Miller 1990]. The primitives to formulate query & descriptions should be in the ontology. INFOMAP serves as the ontology of our QA system. Comparing with the WordNet, the following features: hypernymy, hyponymy, antonymy, semantic relationship, and synset may have a similar counter part in Information Map corresponding to Category, Event, property, and synonym.

2.4 Matching A Natural Language Query

Our kernel program can map a natural language query into a conditioned FAQ and uses the edited knowledge to recognize the concepts in the user’s queries. We use a firing mechanism to propagate nodes in INFOMAP. This mechanism is similar to that used in neural network. When a node is fired, it

will propagate to all the related nodes until some prohibiting signal is generated.

A firing is a labelling action from a node to its parent or to one of its children. For example, if node a fires b, then b is labelled from a. A firing from a node to its parent is called a bottom-up firing. A firing the other way around is called a top-down firing. A bottom-up firing often continues from a node to its parent and then to its grandparent and so on so that a node can fire all of its ancestors if necessary. The purpose of a bottom-up firing is to find the "context" of a given text that is either a sentence or a paragraph. Suppose we want to find the event of the following sentence: "How do I invest in stocks?" and suppose the interrogative word "how" can fire the word "method". Then along the path from "method" to "stock" the above sentence has fired the concepts "stock" and "invest". Hence, the above sentence will correspond to the path:

stock - event - invest - attribute - method

On the other hand, the purpose of a top-down firing is to locate a related concept of interest such as finding a related event or script. INFOMAP provides a mechanism to represent the knowledge that can reflect the number of different ways people may ask a question on the same query concept. First of all, the synonymous expressions for each word concept can fire that word so that it can be substituted in a sentence. Secondly, synonymous events can create sentence-level substitution. Such substitution can cover synonymous expressions of a FAQ. Besides these semantic constraints, the firing mechanism can incorporate various syntactic structures (and templates) plus combinations of both semantic and syntactic structures. Therefore, INFOMAP can be used to parse Chinese sentences provided that enough knowledge about the event structures are given.

A complete description of the INFOMAP would take more than double the size of this paper and shift the focus of our current topic. Hence, we shall only touch upon a few features. Interested reader can refer to [Hsu 1999a].

In INFOMAP, a query concept is usually a path from the root to a node though it could also form a cluster of nodes. The path could consist of a noun (or its synonyms) and a verb or another noun (one of its attributes) or a series of them provided that they form a meaningful event. Since the tree-like knowledge structure can be very deep and very wide, a query concept can also be very deep and wide. If one concept in the hierarchy is fired, all of the hypernymy are fired by propagation since a hypernymy can be a general representation of all of its hyponymy. But this propagation will stop if the hypernymy is a function node, such as "Event" or "Attribute".

Given a natural language query, the system matches the characters and words in the query sentence against node names in the INFOMAP to locate the desired query concept. In general, there is a weighting scheme to select

the most probable query concept. If most of the nodes are located on the path of an FAQ, then we say that the FAQ is fired, i.e., we have matched the open query to this FAQ.

2.5 FAQ Triggered Dialogue System

The dialogue system uses the same knowledge base and can guide the user to query from a shallow concept to the deeper ones. In previous sections we explained how to match a natural language query to a query concept using INFOMAP. In this subsection, we shall demonstrate how to accommodate the knowledge of a dialogue system in INFOMAP.

One can imagine that an "answer" to a high level query could simply be a strategy that intends to guide the user to a deeper query concept. Such a strategy can be context dependent, namely, it can vary based on the user profile and the sequences of questions and answers that this user has just been through in this session. Thus, our INFOMAP has the "dialogue" function nodes to accommodate the implementation of these strategies. Within each dialogue node, there are ask-back questions, each associated with profile conditions and expected actions. The Expected actions intend to let the user focus on predefined actions (such as a selection menu) so that a meaningful dialogue between the user and the system can continue. Each profile condition specifies when this particular ask-back question should be returned and could also be used to filter predefined actions.

A transition diagram can be used to describe each strategy where the nodes in the diagram are the dialogue nodes in our INFOMAP. The activity sequences are controlled by the profile conditions under each dialogue node. If the user digresses from expected actions, then the system will assume the user has initiated a new query and thus, start a new session.

Our INFOMAP currently is designed to represent relatively "simple" query concepts, or in other words, simple events. For questions about the comparison of two distinct query concepts or those consisting of "composite" events (such as events describing the noun of another event), our system decomposes such a complex query into simple query concepts first, and then decipher the relationships among these simple concepts. Since our INFOMAP already contains basic query concepts together with the knowledge about various relationships, the mechanism for the decomposition is simply to design a flow chart to guide a high level query down to more specific ones. In this manner, we can regard our dialogue principle as a process of decomposing a fuzzy query into specific query concepts through a strategy. Traditionally, dialogue model has two categories, the dialogue grammar model and the plan-based dialogue model [Cohen 1996]. Our model contains a dialogue grammar, which has

expectations and can handle exceptions. Thus, it is different form the traditional adjacency pairs.

3. QUESTION ANSWERING SYSTEM REPORT

We build two QA systems, one is for Academia Sinica, a highest national research institute, and the other is for Polaris, a stock trading company. Academia Sinica is a government funded organization and has several thousands of employees. The amount of information in the Web site is very large. The number of Web pages is well over 80,000. It is hard to find deep information from the Web pages of 26 independently running institutes in a unified manner. We have collected 633 distinct FAQs. With the possible combinations of different conditions, the number of query concepts can be as many as 14542. Among them, 8159 query concepts have associated URLs that answers the query. The remaining query concepts have no answer so far (because of the lack of web pages). The number of condition is the number of possible values of the variable in a distinct question. The INFOMAP for these FAQs has 8258 nodes. It takes 10.94 nodes to represent one FAQ on the average. The colleagues of our institute have tested the Academia Sinica QA System, and the data is as follows. We have collected 11802 logs of user queries. There are 3030 different questions, which are mapped by the system onto 924 distinct FAQs. We also find that the top 20% distinct FAQs cover about 80% of the logs.

Table 1: Logs statistical information

Logs	Distinct Open Question	Distinct FAQ
11802	3030	924

The same mechanism has been implemented to build a financial consultant system, “Dr. E”, which is a Q & A System in a web site of Polaris. The service is open to all Internet users and users can query in natural language for various financial information, which is organized by Polaris staff. Financial knowledge is an important asset and organization memory in the firm. It is stored in many internal document and web sites. Polaris collected 6000 FAQs and used INFOMAP to represent the hierarchical knowledge of concept and language. The INFOMAP we build for Polaris has 3394 FAQs and 37148 concept nodes. It takes 13.04 nodes to represent one FAQ on the average.

4. DISCUSSIONS AND CONCLUSIONS

Since it is impossible to record all useful information of an organization, our system only deals with the most frequently asked information. Whenever

it is difficult to gather statistics on FAQs, the map editor has to make a subjective decision. Our process of knowledge acquisition consists of two phases. In the first phase, collect questions whose answers are important to the organization. Then store the questions in INFOMAP. In the second phase, find appropriate answers to the questions and store them in web pages.

Our INFOMAP consists of two types of knowledge: linguistic knowledge and domain knowledge. Our FAQ-Centered dialogue system has been applied to customer service and web CRM. Future research is to design abstract conceptual script to extract the desired knowledge in INFOMAP automatically. The adoption of an abstract semantic map and concrete sample maps will become increasingly important.

References

[Burke et al. 1997] R. Burke, K. Hammond, V. Kulukin, S. Lytinen, N. Tomuro and S. Schoenberg. Experiences with the FAQ system. AI Magazine, 18(2):pp57-66, 1997.

[Cohen 1996] Phil Cohen, Dialogue Modeling, in A Survey of the State of the Art in Human Language Technology, section 6.3, Eds. Ron Cole, 1996.

[Guarino 1999] Guarino N., Masolo C., and Vetere G., OntoSeek: Content-Based Access to the Web, IEEE Intelligent Systems 14(3) , pp. 70-80, May/June 1999.

[Hsu 1999] W. L. Hsu and Yi-Shiou Chen, On Phoneme-to-Character Conversion Systems in Chinese Processing, Journal of Chinese Institute of Engineers 5, (1999), 573-579.

[Hsu 2001] Wen-Lian Hsu and Shih-Hung Wu, Event Identification Based On The Information Map - INFOMAP, to appear in symposium Natural Language Processing and Knowledge Engineering of the IEEE Systems, Man, and Cybernetics Conference, 2001.

[Jocobs 1993]R. S. Jacobs and L. F. Rau. Innovations in text interpretation. Artificial Intelligence, 63(1-2):143-191, 1993.

[Kupiec 1999] Julian M. Kupiec, MURAX: Finding and Organizing Answers from Text Search, in Tomek Strzalkowski, editor. Natural Language Information Retrieval. Kluwer Academic Publishers, 1999.

[McCune 1985]B. McCune, R. Tong, J.S. Dean, and D. Shapiro. Rubric: A system for rule-based information retrieval. IEEE Transaction on Software Engineering, 11(9), 1985.

[Miller et al. 1990] Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine J. Miller**.** "Introduction to WordNet: an on-line lexical database." In: *International Journal of Lexicography* 3 (4), pp. 235 - 244. 1990.

[Staab 2001] Staab, S., H.-P. Schnurr, R. Studer, Y. Sure: Knowledge Processes and Ontologies. IEEE Intelligent Systems, Special Issue on Knowledge Management, 16(1), January/February 2001.

[Stein 1995] E. W. Stein, Organizational memory: Review of Concepts and Recommendations for Management. *International Journal of information Management*. Vol. 15. No2, pp.17-32, 1995.

[Strzalkow1999] Tomek Strzalkowski, editor. Natural Language Information Retrieval. Kluwer Academic Publishers, 1999.

[Winiwarter 2000] W. Winiwarter, Adaptive natural language interfaces to FAQ knowledge bases, *Data & Knowledge Engineering*, 35, (2000), 181-199.