EVENT IDENTIFICATION BASED ON THE INFORMATION MAP - INFOMAP

Wen-Lian Hsu, Shih-Hung Wu and Yi-Shiou Chen

Institute of Information Science, Academia Sinica
Nankang, Taipei, Taiwan, R.O.C.
hsu@iis.sinica.edu.tw, shwu@iis.sinica.edu.tw

Abstract

We present a knowledge representation scheme, INFOMAP, together with a mechanism that matches the event of a natural language sentence with part of the domain ontology in the INFOMAP. The design of this scheme is to facilitate both human browsing and computer processing of the domain ontology. INFOMAP is also a knowledge framework designed to facilitate knowledge sharing by different application systems. We constructed a question answering system to demonstrate the power of INFOMAP. When the QA-system receives a user's query, it will extract the corresponding events or scripts based on the ontology in INFOMAP. The understanding of a question involves extracting such information as the question type, the question subject, the question condition and the question context. A dialogue on the question is triggered at the same time to guide the user to retrieve more relevant information.

Keywords
Information Map, Ontology, Question answering system, Dialogue.

1. Introduction

We design a knowledge representation scheme, INFOMAP to facilitate both human browsing and computer processing of the domain ontology in the system. The domain ontology is constructed from structured concepts in each specific domain. Examples of concept structures range from simple concepts such as a word, a phrase, or an event to more complex ones such as a sentence, a paragraph, a script (a collection of related events), a story, or the passive tense of English and etc. Each concept is associated with a structure (a sub-map) describing the relationships of this concept with its related concepts. The system can store and generate a large amount of events, syntactic or semantic structures and scripts. Given a natural language sentence, the system tries to match it to a sub-map or decompose it into several events in INFOMAP. An event consists of a sequence of relevant nouns or a pair of relevant noun and verb, which represents a topic noun and some modifiers.

With this concept-based INFOMAP, domain knowledge storage and retrieval can be carried out at the same time. In other words, whenever a new piece of knowledge is added to the map, its utilization scheme (such as retrieval) is automatically defined and can be carried out by the computer.

Enormous amount of knowledge need to be constructed in the design of large-scale natural language application systems such as speech recognition [8], natural language agents [9], organizational memory [21], machine translation, text-to-speech and grammar-check systems. It is often the case that similar knowledge content may need to be reproduced in different ways to facilitate different application systems. INFOMAP is a knowledge framework designed to facilitate knowledge sharing among different application systems. Previous works on knowledge representation schemes, including predicate logic, frames [14], conceptual graphs [19], semantic nets [16] and scripts [17], have many useful features. Our concept-based approach can incorporate most of their features in a consistent and versatile framework.

As an application, we constructed a question answering system based on INFOMAP. When the QA-system receives a user's query, it attempts to extract the corresponding events or scripts from the domain ontology in INFOMAP. The understanding of a question involves extracting such information as the question type, the question subject, the question condition and the question context. The QA-system makes use of this information to determine whether the question can be correctly answered.

Most of the information retrieval works focus on extensive information domains such as the World Wide Web, where the demand for precision is not terribly strong. In this situation, keyword search is both efficient and practical. However, for intranet services where a more detailed information and right-to-the-point link (or answer) is adequate, the techniques of regular search engines are far from desirable. For example, imagine a user is seeking for information services in a government organization (such as getting a new license plate). Since most users are not familiar with the correct keywords, they need to browse a lot of pages to "identify" the jargons to use. In another situation, popular keywords are ubiquitous in different pages and the user needs to specify the relations among the keywords to identify the most relevant pages (or paragraphs or even sentences). Any keyword-based search engine would not be very helpful. Although the techniques of bi-gram or tri-gram are somewhat useful, they still could not capture the relations of keywords. Instead, in our knowledge representation scheme, the structures of keywords will be naturally produced to facilitate the search. This will be elaborated in Section 2.

To tackle the situation where a user does not know how to form a detailed query, a simple-minded question answering mechanism is barely enough. Rather, an interactive dialogue between the user and the system is more user-friendly and effective. In Section 3, we shall discuss how to incorporate a dialogue system into an existing QA system.

1.1 Related work

We shall discuss several related QA systems and dialogue systems. The Murax system [11] determines from the syntax of a question if the user is asking for a person, place, or date. It then tries to find sentences within encyclopedia 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 FAQ matching system [4] is to match question-style queries against question-answer pairs. The system may use some standard IR search 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. Ask Jeeves, an Internet service provider, applies a less automated approach to question answering [1]. Human selected Web sites are first matched to a predefined set of question types. The question types are then matched against user's natural language query. Then user selects the most accurate rephrase of the query and the reformulated query is linked to suggested Web sites. There are some systems [10,12,20] attempt to parse natural language queries in order to extract concepts to match against concepts in the text collection. In some systems, texts and questions are uniformly processed and semantic paths between concepts structured around WordNet [13] are established using a marker propagation method [7].

Dialogue is a notable user interface in modern information retrieval [2]. In their paper, dialogue is used to help the user to form a retrieval process by dialogue on the query concept. Dialogue-based interfaces have been explored in information retrieval research to mimic the interaction between the user and a librarian. The THOMAS system [15] provided a question and answer session within a command-line-based interface. Belkin et al. [3] defined an elaborate dialog interaction models.

2. Information Map

In this section, we shall introduce our knowledge representation scheme that matches a natural language query to a query concept. We shall define the question portion of a FAQ question-answer pair to be a *query concept*. A query concept may contain several basic concepts in the structure. In order to respond appropriately, the system needs the knowledge about how many different ways people may ask a question on the same query concept.

2.1 Query representation

INFOMAP has a tree-like structure though this is only a deceivingly simple statement since it does contain "references" that connect nodes on different branches. The root node is usually the name of a domain or a subject such as passport or 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.

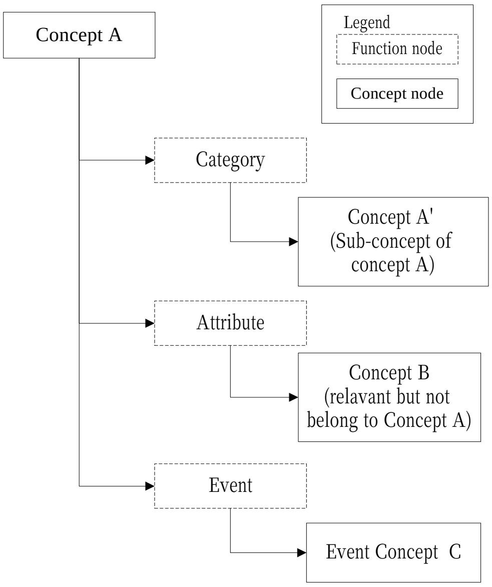

Concept A
Legend
Function node
Concept node
Category
Concept A' (Sub-concept of concept A)
Attribute
Concept B (relavant but not belong to Concept A)
Event
Event Concept C

Figure 1. Component of the INFOMAP

Figure 1 shows the basic component of

INFOMAP. Where Concept *A* is a topic, and a sub-tree of the taxonomy hierarchy of concept *A* is under the function node category. This is very similar to the directory structure in most Internet yellow page. Concepts that are relevant to concept *A* but not belong to concept *A* are listed under the function node attribute. Associated events, i.e., actions that can be associated with concept *A*, are listed under the function node event. For example, if concept *A* is a "car", then it can be driven, parked, raced, washed, repaired and etc.

2.1.1 Function nodes

There are some nodes, called function nodes, to label the relation between two other nodes in INFOMAP. The basic function nodes are: category, attribute, example, synonyms and event. There are some function nodes to build the QA system, such as FAQ, FAQ condition, test query and some other infrequently used function nodes. These function nodes help to represent and identify query concepts. The synonym of a concept is listed under the function node synonym of each concept. Under the function node "example" are the examples of a concept. For example, if the concept is "hotel", then its examples can be the actual hotel names. Function node "FAQ" gives a typical question associated with the concept. Function node "FAQ conditions" are items in a query that can be substituted by examples such as cities, hotels and etc. There are also other function nodes that are used specifically for linguistic feature extraction and parsing. Function node "grammatical constraints" check the syntactic constraints on the concept in a phrase or sentence.

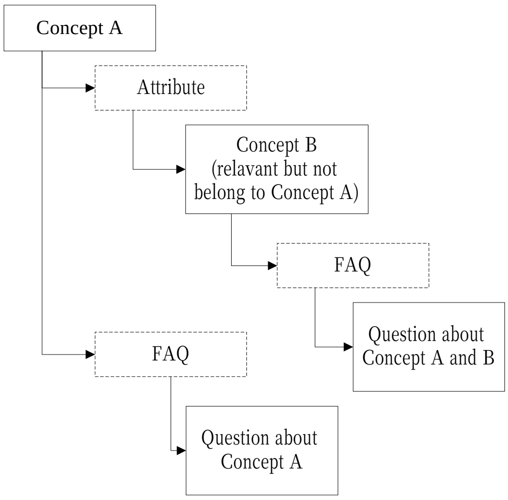


Figure 2. FAQ and query concepts

Figure 2 shows two FAQs and the associated query concepts. The FAQ under concept *A* is a question about concept *A* only. The FAQ under concept *B* is a question about concept *A* and *B*. Therefore, we can distinguish different questions according to the concepts in these questions. The concept *A* and *B* are also used as a way to match an open query into the FAQ. Some other concepts that are relevant to the FAQ can be added under the FAQ nodes, served as addition information, such as the interrogative and FAQ condition and FAQ test query.

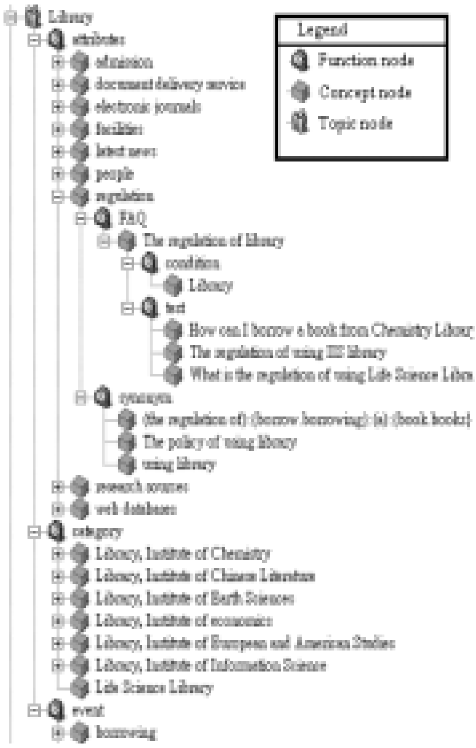

Library
attributes
admission
document delivery service
electronic journals
facilities
latest news
people
regulation
FAQ
The regulation of library
condition
Library
test
How can I borrow a book from Chemistry Library
The regulation of using IIS library
What is the regulation of using Life Science Libra
synonym
(the regulation of) (borrow borrowing) (a) (book books)
The policy of using library
using library
research sources
web databases
category
Library, Institute of Chemistry
Library, Institute of Chinese Literature
Library, Institute of Earth Sciences
Library, Institute of economics
Library, Institute of European and American Studies
Library, Institute of Information Science
Life Science Library
event
borrowing
Legend
Function node
Concept node
Topic node

Figure 3. A partial view of the ontology about the libraries in Academia Sinica

Figure 3 shows the related information of library (in Academia Sinica) and some FAQs. It can be seen from the figure that the information of a library is divided into three functions nodes: event, category 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 sciences 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 and so on. The event of library can be "borrowing books from the library". Nodes under the "Category" form a hierarchy, which is taxonomy of library. Each node under "attribute" forms a new hierarchy, which is not part of the taxonomy hierarchy of library, but is relevant to it. The nodes under “event” are similar. We classify the relevant concept into “attribute” and “event” based on the concept belong to noun or verb. This is a special criterion of Chinese, since the verb and noun have no morphological difference in Chinese.

2.1.2 INFOMAP serving as the Ontology

Ontology represents domain knowledge and serves as a common understanding of the domain. Ontology consists of definitions of concepts, relations and axioms [6,18]. 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 [13]. The primitives to formulate query and 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 INFOMAP corresponding to Category, Event, attribute, and synonym.

2.2 Query understanding

A query concept is usually a path from the root to a node, though it could also form a cluster of nodes in the INFOMAP. 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. 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.

The example in Figure 2 illustrates how the function node “condition” can make FAQ more flexible. Considering the FAQ “The regulations of library” under the node “regulation”, there is a condition node “library” in the figure. This condition refers to any specific library under the category of library. Thus, the system can recognize the question: “What is the regulation of Life Science Library?” The function node “condition” acts like a variable to represent these libraries in related questions.

2.2.1 The firing mechanism

In order to understand the meaning of an open query, we designed a firing mechanism that can identify the most probable context and the most likely FAQ.

First of all, the open query will fire the nodes in the INFOMAP in a rough way. The words in the open query will fire the corresponding nodes. These fired nodes can help to determine the possible context. We call this phase “topic speculation”. This is a scoring mechanism that calculates the distribution of the nodes in each topic context, which is called agent, and then finds the possible target agents. Next, firing all the nodes in the target topics in a detailed way. All the nodes in the target topics will be the candidate of template firing, propagation firing and reference firing. Once the firing process ends, then the system collects the nodes in the context. A fired node in the target topics will be assigned a score according to the length of the string that fired this node. Afterward, collecting FAQ nodes in the target topics, and calculating a score for each FAQ according to the nodes on the path from root to this FAQ node. Finally, sort the FAQ according to the total score.

2.2.2 The scoring mechanism

The scoring mechanism is to rank those FAQs that are more similar to the open query in the context.

The score of a fired node is proportioned to the string that fired this node. If this node is in a predefined “non-context” area, then its score is one per Chinese character. Otherwise, its score is ten per Chinese character. The non-context area contains nodes, which represent common concepts. These concepts are less helpful to identify the context of an open query.

The context score of a FAQ is the total score on the path from the root to the FAQ node. The ancient nodes that are fired should have scores. The total context score of a FAQ is the summation of agent speculation score, context score and specified context score. Where the specified context score is the score of the referenced nodes that are relevant to this FAQ. The FAQs in the context are sorted according to the total context score. In the case that the total context scores of two FAQs are the same, we will consider if they are on the same path. If they are on the same path, then the deeper one is preferred.

2.3 The Academia Sinica QA system

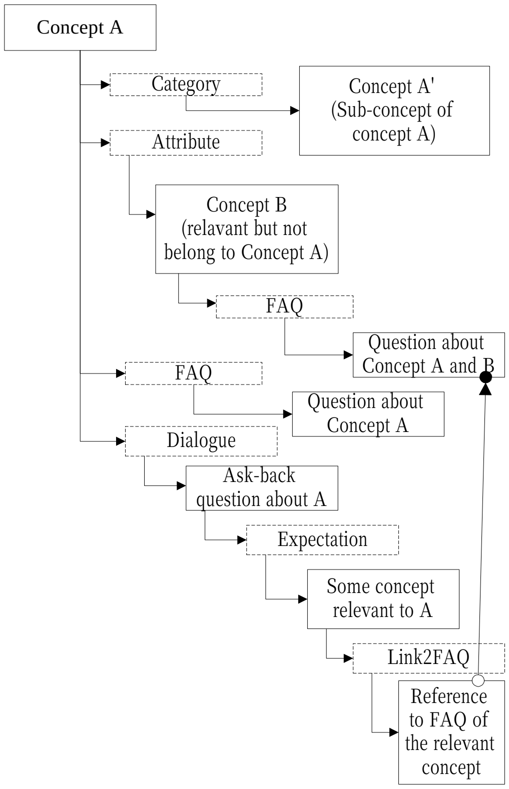
We have implemented a QA system using the ontology and offered service online for Academia Sinica. Academia Sinica is a government funded research organization, which has 26 independently running institutes and several thousand employees. The amount of information in the Websites is very large. The total number of Web pages is well over 80,000. It is hard to find information from the Web pages of 26 institutes in a uniform manner. Therefore, we constructed the Academia Sinica QA System to retrieval information from them. We collected and identified 626 distinct FAQ question types. With the possible combination of different conditions, the number of query concepts can be as much as 12,876. Among them, 10100 query concepts have associated URLs that answers the query. The remaining query concepts have no answer yet.

3. The FAQ-Centered Dialogue System

In the previous section we explained how to match a natural language query to a query concept using INFOMAP. In this section, we will extend the INFOMAP to do the dialogue by adding more function nodes.

3.1 FAQ-triggered dialogue

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 varies based on the user profile and the sequences of questions and answers that this user has just gone through in this particular session. Thus, our INFOMAP has a “dialogue” function node to accommodate the implementation of the strategies.



Concept A
Category
Concept A' (Sub-concept of concept A)
Attribute
Concept B (relavant but not belong to Concept A)
FAQ
Question about Concept A and B
FAQ
Question about Concept A
Dialogue
Ask-back question about A
Expectation
Some concept relevant to A
Link2FAQ
Reference to FAQ of the relevant concept

Figure 4. Dialogue and relevant concept

Any FAQ of concept *A* may have an associated dialogue function node. Which means that once the FAQ about concept *A* is fired, then a dialogue on this FAQ can be triggered. Under each dialogue node, there are ask-back questions, each associated with profile conditions and expected actions. Each profile condition specifies when this particular ask-back question should be returned. 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. The expected answer of the ask-back question is under the function node “Expectation”. Once the expectation is satisfied, it will link to another FAQ. In figure 4, the FAQ about concept *A* has an associated dialogue. And there is an ask-back question about concept *A*. If the user’s reply is within the expectation, then the dialogue system will guide the user to a deeper concept. In this example, lead to a FAQ about concept *A* and *B*.

If the user digresses from expected actions, then the system will assume the user has initiated a new query and start a new session.

3.2 The dialogue principle

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 to decipher the relationship among these simple concepts. According to the cooperative principle [5], we believe that the two parties in a dialogue will work together to achieve a better mutual understanding. Therefore, in the dialogue system, there are 6 possible roles for each action. These are: query, answer, ask-back question, expect answer, expect query, exception. The interaction is richer than that of a naïve QA system where interaction is a sequence of question and answer, that is, each action must be in one of these two roles.

4. Conclusions

The purpose of INFOMAP is to identify the possible topic and event in a sentence. The INFOMAP knowledge structure is a robust way to represent the knowledge; therefore, several knowledge editors may work together. We also provide a distributed interface that many knowledge editors can work together easily.

With the ontology we built in INFOMAP, it is possible to implement many applications, such as QA system and dialogue system. Since human dialogue is a natural way to exchange information between people. There is much implication information helping people to understand each other during a conversation. This knowledge can be represented in the INFOMAP. Then, the dialogue, as a human-computer interface, can help the domain expert (knowledge editor) to communicate with the end users asynchronously. In our INFOMAP, the set of the function nodes determines a certain class of query concept. To build a real world system, we have used our INFOMAP to represent a set of 612 distinct FAQ of a specific domain. Identifying more useful function nodes can certainly improve the capability of our INFOMAP as well as FAQ-centered dialogue system.

Reference

1. Ask Jeeves: http://www.askjeeves.com, 1998.
2. R. Baeza-Yates, B. Ribeiro-Neto, Modern information retrieval, Addison-Wesley, 1999.
3. N. Belkin, P. G. Marchetti, and C. Cool. Barque – design of an interface to support user interaction in information retrieval. Information Processing and Management, 29(3):325-344, 1993.
4. 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.
5. H.P. Grice, Logic and conversation, in Speech acts, edited by Peter Cole and Jerry L. Morgan, New York : Academic Press, 1975.
6. 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.
7. S. Harabagiu and D. Moldovan, An Intelligent System for Question Answering, in the Proceedings of the 5th Conference on Intelligent Systems, Reno NV, pages 71-75, 1996.
8. Wen-Lian Hsu and Yi-Shiou Chen, On Phoneme-to-Character Conversion Systems in Chinese Processing, Journal of Chinese Institute of Engineers 5, (1999), 573-579.
9. Wen-Lian Hsu, Yi-Shiou Chen and Yuan-Kai Wang (1999), Natural language agents - An agent society on the Internet, Proceedings of PRIMA 99.
10. R. S. Jacobs and L. F. Rau. Innovations in text interpretation. Artificial Intelligence, 63(1-2):143-191, 1993.
11. Julian M. Kupiec, MURAX: Finding and Organizing Answers from Text Search, in Tomek Strzalkowski, editor. Natural Language Information Retrieval. Kluwer Academic Publishers, 1999.
12. 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.
13. 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.
14. Marvin Minsky, A Framework for Representing Knowledge, Reprinted in The Psychology of Computer Vision, P. Winston (Ed.), McGraw-Hill, 1975.
15. R. N. Oddy. Information retrieval through man-machine dialogue. Journal of Documentation, 33:1-14, 1977.
16. M.R. Quillian, Semantic Memory, In M. Minsky (ed.), Semantic Information Processing, pp. 227-270. Cambridge, MA: MIT Press, 1968.
17. Schank, R.C. and R. Abelson, "Scripts, Plans, Goals and Understanding", Hillsdale, NJ: Lawrence Erlbaum, 1977.
18. 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.
19. Sowa, John F., Conceptual graphs for a database interface, IBM Journal of Research and Development, vol. 20, no. 4, pp. 336-357, 1976.
20. Tomek Strzalkowski, editor. Natural Language Information Retrieval. Kluwer Academic Publishers, 1999.
21. Shih-Hung Wu and Wen-Lian Hsu, FAQ-centered Organizational Memory, to appear in the KM/OM workshop of IJCAI-01, Seattle, U.S.A, 2001.