UTILITY OF WEAKLY STRUCTURED MEMORY ORGANIZATION FOR INTEGRATING HETEROGENEOUS INFORMATION

HARUMI MAEDA KAZUTO KOUJITANI[†] TOYOAKI NISHIDA

Graduate School of Information Science,
Nara Institute of Science and Technology,
8916-5,Takayama,Ikoma,Nara 630-01 JAPAN
Phone: +81-7437-2-5265 Fax: +81-7437-2-5269
E-mail: harumi-m@is.aist-nara.ac.jp

Abstract

This paper proposes the use of *weakly structured memory organization* as a means of helping people develop and maintain memory from heterogeneous information sources. In order to endorse the claim, we have implemented a system called CM-2. CM-2 is characterized as an *intelligent amplifier* whose primary design goal is enhancement of human intelligence. CM-2 helps users: (a) build the memory containing multimedia information, (b) organize the memory from multiple perspectives, (c) incorporate and integrate information from external sources (e.g., WWW, CD-ROM), (d) access the memory in an associative fashion, and (e) incrementally refine and organize the memory. We report the evaluation of CM-2 against two test cases: (1) ontology development and (2) information reorganization of WWW pages.

keywords: weakly structured memory organization, ontology development, information reorganization

1 Introduction

The purpose of this research is to help people in the information age develop and maintain memory (either personal or group) from diverse heterogeneous information sources from the world-wide information network. Critical issues there are: (a) information gathering and filtering, (b) information classification and evaluation, (c) handling multimedia information and media conversion, (d) incremental and adaptive construction of the memory structure, and (e) information retrieval from the memory structure. This paper addresses the last two problems.

In this paper, we point out that use of *weakly structured memory organization* is effective in achieving our goal and endorse the claim by an implemented system called CM-2 and a couple of experiments. CM-2 is rather characterized as an *intelligent amplifier* whose primary design goal is enhancement of human intelligence, rather than an artificial intelligence which exhibits intelligence behavior by itself. In other words, our research is characterized more as a knowledge media research[1] than as an AI research.

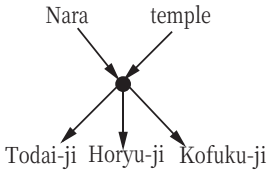


Figure 1: Weakly Structured Memory Organization

The technical contribution of this paper is the weakly structured memory organization consisting of *associations*, each of which is a many-to-many mapping of memory units. In a couple of test cases, we have made a preliminary experiments on generating associations from existing HTML documents and refining them.

In what follows, we first describe the role of weakly structured memory organization and overview the CM-2 information base system. We then report the evaluation of CM-2 against two test cases: (1) ontology development and (2) information reorganization of WWW pages.

2 Weakly Structured Memory Organization

We introduce *Weakly Structured Memory Organization* as a primary information structure for articulating conceptually diverse information by aggregating conceptually relevant information. The basic entities of weakly structured memory organization are (a) a *unit* which represents either a concept or an external datum, and (b) an *association* which connects a collec-

[†]Presently with OMRON Corporation

tion of key concepts with a collection of units which is normally reminded by the given keys. Figure 1 shows an example of associations. This denotes that from given concepts "Nara" and "temple", one may be reminded of "Todai-ji", "Horyu-ji", and "Kofuku-ji".

3 CM-2 Information Base System

CM-2[1] is a knowledge media information base system on which humans and computers collaborate each other to accumulate, share and explore diverse information gathered from heterogeneous information sources.

CM-2 helps the user or the group of users: (a) build the memory containing multimedia information, (b) organize the memory from multiple perspectives, (c) incorporate and integrate information from external sources (e.g., WWW, CD-ROM), (d) access the memory in an associative fashion, and (e) incrementally refine and organize the memory.

In addition to basic editing, browsing and retrieval mechanisms, CM-2 has four facilities to support these user activities and they are stated below.

Information Capture Facility Information capture facility generates associations from various information sources. The general procedure of information capture facility for WWW pages consists of the following steps: (1) extraction of noun phrases and generation of units based on morphological analysis[2] and (2) generation of associations by analyzing the structure of HTML documents.

Information Refinement Facility Information refinement facility refines incoherent associations into coherent associations using heuristics and measurement of similarity.

IS-A Relation Generation Facility IS-A relation generation facility generates IS-A relations by analyzing the class of given units using heuristics.

Frame Generation Facility Frame generation facility generates frames, in other words, reorganizes given associations to a set of entities and attributes by path-finding.

4 Test Cases

4.1 Test Case 1: Ontology Development

Ontology plays a central role in achieving large scale knowledge sharing. Unfortunately, development of ontology is often a quite painstaking and time consuming task, because it needs much effort to collect and select terms through task analysis. We apply our approach based on the weakly structured memory organization to ontology development. It allows for data-driven ontology development, by accumulating raw data and incrementally creating the structure of concepts through human-computer collaboration (raw data → associations → refined associations → semantic networks → frames → predicate logic) . The overall process of our approach is shown in Figure 2.

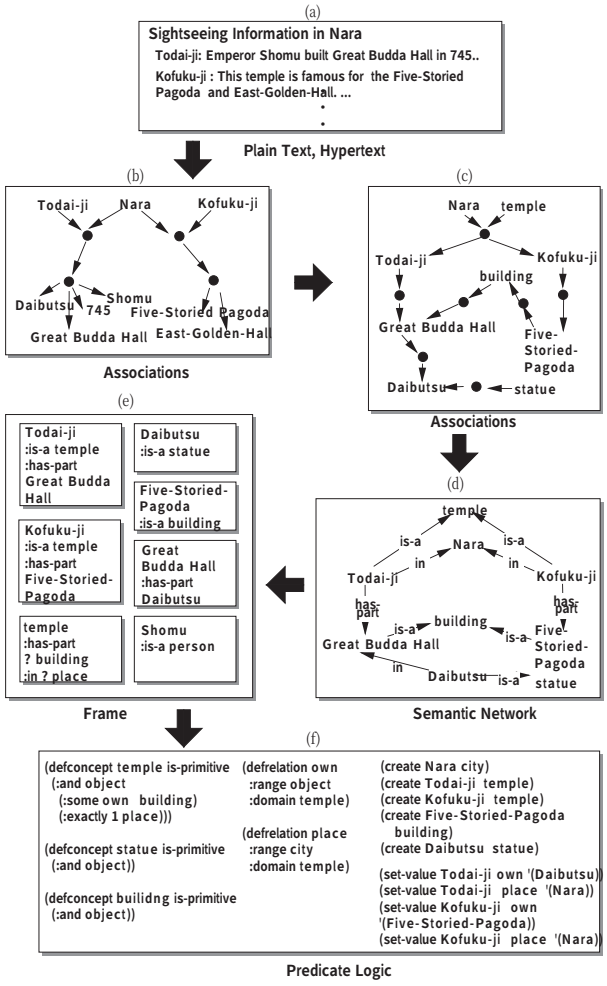


Figure 2: Test Case 1: Ontology Development

Experiment We gave 30 WWW pages concerning ARPA Intelligent Integration of Information (I3) Initiative [2] to CM-2. Each page contains overview of projects which belong to the I3 Initiative. CM-2 generated associations by information capture facility. Generated associations themselves are incoherent and cannot be used as ontology as they are. Information refinement facility assisted users to formate these associations into more coherent structure. Figure 3 shows

[1] "CM" stands for "Contextual Media" which stands for our long term theoretical goal.

[2] http://dc.isx.com/I3/

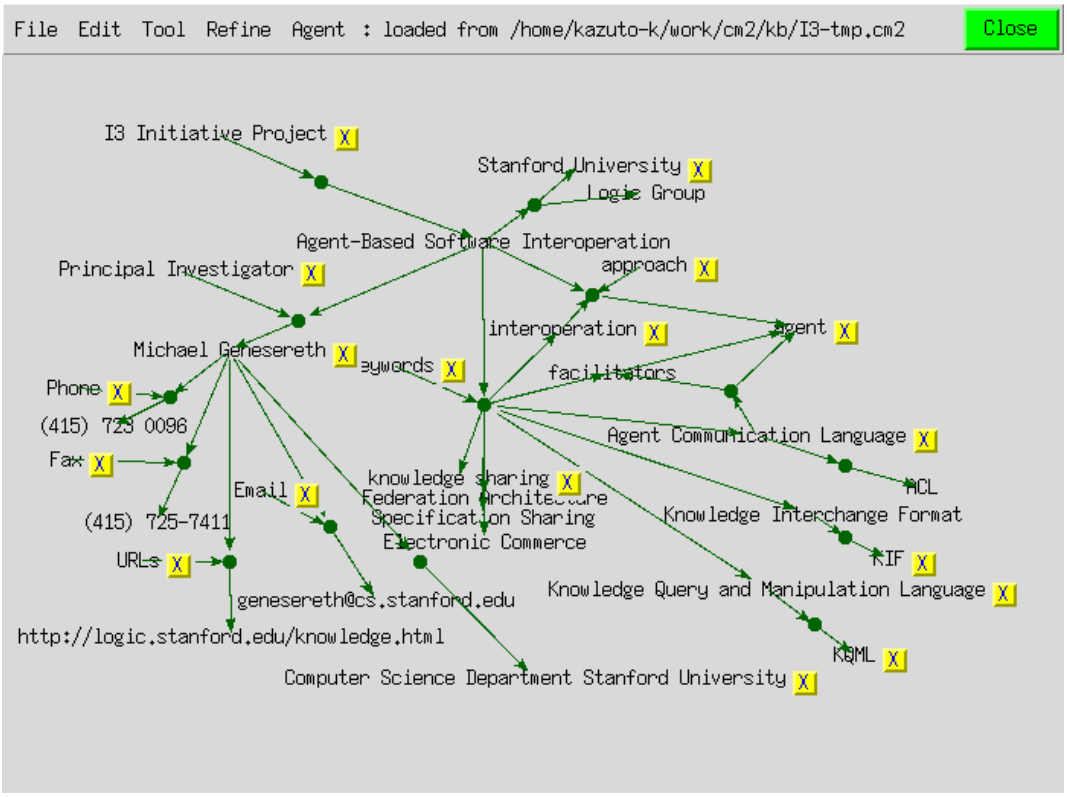


Figure 3: An Example Result of Refined Associations

an example result of refined associations concerning the project "Agent-Based Software Interoperation". It corresponds to level c ontology (refined associations) in Figure 2.

4.2 Test Case 2: Information Reorganization of WWW Pages

It is difficult to reorganize relevant information from large-scale diverse information on World Wide Web. To integrate a wide variety of diverse information on WWW, we use the weakly structured memory organization to extract information from WWW pages and reorganize it according to user's input. The overview of the process is shown in Figure 4.

Experiment We gave 100 WWW pages concerning AI researchers to CM-2 for organizing AI directories. CM-2 extracted units about 7 classes (researchers, projects, e-mail, topics, universities, departments and laboratories), and generated associations. CM-2 reorganized these associations to display various directories according to user's input.

Figure 4 (a) shows an example result when a user inputs "reasoning","researcher","e-mail","project" and "university" and selects "table". For instance, researchers such as Alon Levy, Edward Feigenbaum and James Allen are extracted first, because the word "reasoning" and their names are written near in WWW pages, and then their related information is reorganized.

Figure 4 (b) shows an example result when a user inputs "project", "researcher","e-mail","university" and selects "list". In this case, projects are extracted first and then other relevant information are reorganized.

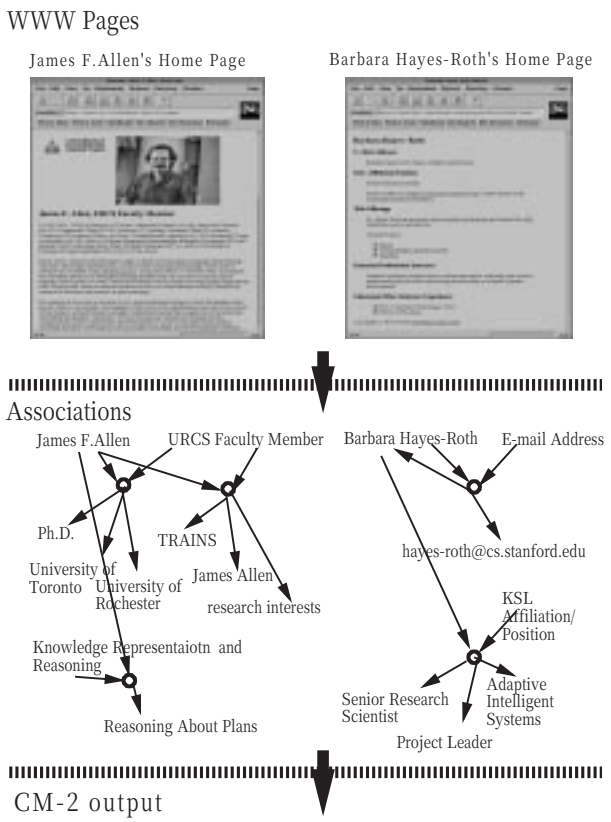
A summary of the results of two tests are shown in Table 1: (1) to display researchers' table (Test 1) and (2) to display projects' list (Test 2). The results of Test 1 (90% at precision rate, 83% at recall rate) are better than those of Test2 (68% at precision rate, 73% at recall rate), because original WWW pages are persons' pages.

Table 1: Result of Test Case 2

Test	Precision	Recall
Test 1 (researcher)	90%	83%
Test 2 (project)	68%	73%

Precision: $\frac{appropriate\ units}{generated\ units} \times 100\ (\%)$

Recall: $\frac{appropriate\ units}{units\ which\ should\ be\ extracted} \times 100\ (\%)$



(a) Test 1 (researcher)

(b) Test 2 (project)

Figure 4: Test Case 2: Information Reorganization of WWW Pages

5 Related Work

Our work is related to recent work on information extraction from heterogeneous sources on the Internet ([3],[4],[5],[6],[7]). Instead of focusing on the strategies and heuristics for information gathering, we concentrate on how to classify information obtained from multiple information sources and integrate it into personal information base. Our approach is to provide a framework of collaborations for human information sharing with a low structural facilities.

6 Conclusions

We proposed the weakly structured memory organization called *associations* as a means of helping people develop and maintain memory from diverse heterogeneous information sources. We described a system called CM-2 which implements our claim. We applied CM-2 to two test cases and the results indicate that our approach is effective for (1) ontology development and (2) information reorganization of WWW Pages.

As a future research, we plan to integrate associations generated by other sources such as Usenet articles.

References

[1] Mark Stefik. The next knowledge medium. *AI Magazine*, 7(1):34–46, 1986.

[2] Eric Brill. Some advance in transformation-based part of speech tagging. In *Proceedings of the Twelefth National Conference on Artificial Intelligence (AAAI-94)*, 1994.

[3] Alon Y. Levy, Yehoshua Sagiv, and Divesh Srivasava. Towards efficient information gathering agents. In *Working Notes of the AAAI Spring Symposium on Software Agents*, pages 64–70, 1994.

[4] Robert Armstrong, , Dayne Freitag, Thorsten Joachims, and Tom Mitchell. WebWatcher: A learning apprentice for the World Wide Web. In *Working Notes of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, pages 6–12, 1995.

[5] Marko Balabanovi'c and Yoav Shoham. Learning information retrieval agents: Experiments with automated web browsing. In *Working Notes of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, pages 13–18, 1995.

[6] Wen-Syan Li. Knowledge gathering and matching in heterogeneous databases. In *Working Notes of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, pages 116–121, 1995.

[7] Michiaki Iwazume, Hideaki Takeda, and Toyoaki Nishida. Ontology-based information gathering and text categorization from the internet. In *Proceedings of the Ninth International Conference in Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, 1996. 305–314.