An Information Representation for Concept Formation

Harumi Maeda, Kazuto Koujitani, Takashi Hirata and Toyoaki Nishida

Graduate School of Information Science,
Nara Institute of Science and Technology

Address: 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: We propose an information representation called associations for concept formation. We set a hypothesis that linking information without defining the semantics using associations is effective to deal with a large amount of diverse information. In our approach, we gather raw information from heterogeneous information sources using associations and gradually form concepts through collaborations between humans and computers. To endorse the claim, we implemented a system called CoMeMo which helps people to construct and share memory in a group or in a community. We report two test cases: (1) integration of creative thinking and information retrieval and (2) ontology development from WWW pages. We show experimental results and make discussion.
Keywords: concept formation, associations

1. Introduction

This paper describes an approach to develop a computational model which facilitates concept formation. Nonaka and Takeuchi [1] analyzed the process of creating corporate knowledge in successful companies in Japan and pointed out the importance of knowledge conversion between tacit and explicit knowledge. They proposed a knowledge spiral model consisting of externalization (tacit knowledge -> explicit knowledge), combination (explicit knowledge -> explicit knowledge), internalization (explicit knowledge -> tacit knowledge), and socialization (tacit knowledge -> tacit knowledge). As a first step, we focus on integrating heterogeneous information in the first two steps.

The research interest behind this approach consists in search for a primary concept structure which plays a critical role for integrating heterogeneous information. We already know the power of knowledge representation languages such as first-order logic or frames which allow computer programs to undertake semantic processing. Unfortunately, classic knowledge representation languages are so logically rigid that one must spend tremendous amount of efforts on knowledge acquisition. This defect

forces too much on human effort and hence significantly hinders accumulation of a large amount of useful information.

The other extreme is doing without concept structure, or at most with very simple structure. At a glance, this approach does not appear to work. However, such a prejudice is not always right. In Japan, a recent best sold book[2] successfully convinced people that organizing information pieces in a chronological order is a very promising way of maintaining personal files. Lifestreams[3] was designed along with most the same idea. The magic is that they effectively make use of the user's own memory to supplement incompleteness of system's memory. Social filtering is based on the similar idea. Good news with this approach is the computational cost for gathering information is low. Of course, there would be an obvious limitation if we were solely dependent on human's ability.

In this paper, we investigate an information representation called *associations* which enables maximal computational support with minimal human effort. We set a hypothesis that linking information without defining the semantics is effective to deal with a large amount of diverse information. Associations are one of such information representations we are investigating; each of which is a many-to-many mapping of information units. In our approach, we gather raw information from heterogeneous information sources using associations and gradually form concepts through collaborations between humans and computers.

Most of information people produce and manipulate (such as natural language documents, or images) is conceptually diverse in the sense that its semantics is not rigorously defined or completely agreed upon. We can think of various stages on which tacit information is turned into well-defined rigorous information. At early stages, keywords or phrases are simply enumerated or arranged into a space, while later they will be replaced by more rigorous expressions and connected by semantically well-defined logical relations. To endorse our claim, we have implemented a system called CoMeMo which helps people to construct and share memory in a group or in a community. In a couple of test cases, we have made preliminary experiments on generating associations from existing HTML documents and integrating them.

In what follows, we first describe the role of the associations and overview CoMeMo. We then report the evaluation of CoMeMo against two test cases: (1) integration of creative thinking and information retrieval and (2) ontology development from WWW pages. Finally, we show experimental results and make discussion.

2. Associations

Associations connect a collection of key *units* (hereafter *keys*) with a collection of units (hereafter *values*) which are normally reminded by the given keys. Units are basic entities of associations which represent concepts, texts, or image files. Example associations are shown in Figure 1.

We call a set of associations collected from a particular point of view *workspaces*. The CoMeMo information base is a collection of workspaces.

In our approach, the semantics of the associations is not defined rigorously. Instead, we leave the interpretation of the semantics to tacit human background knowledge. This facilitates information acquisition from a variety of data (e.g. images, texts, ideas)

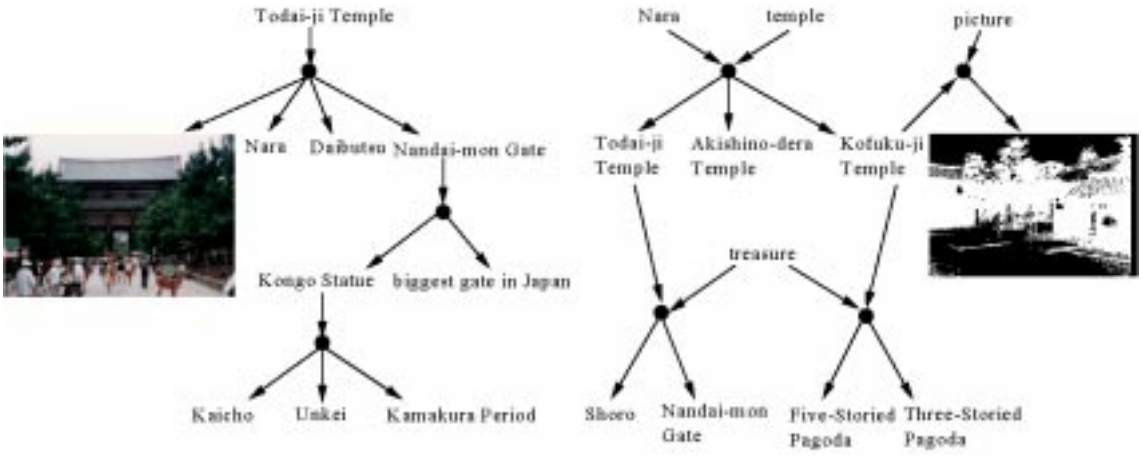

Figure 1. Example Associations

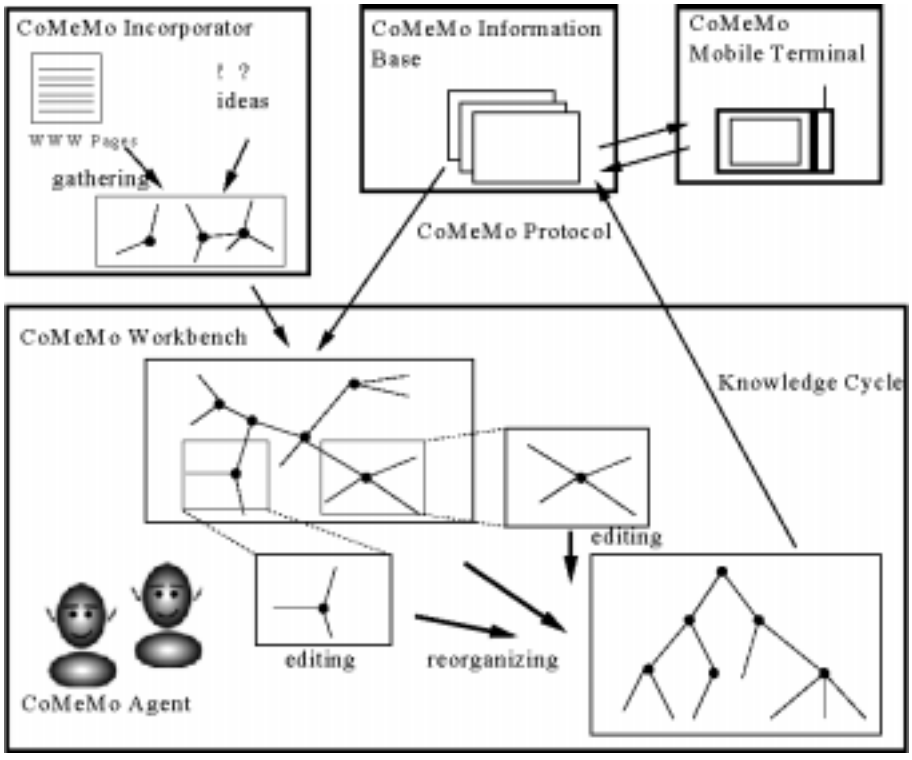

Figure 2. CoMeMo Architecture

3. CoMeMo

CoMeMo is a system which helps people to construct and share memory in a group or in a community.

The CoMeMo architecture consists of following components: (a) **CoMeMo Incorporator** which gathers and converts a wide variety of information into information representations which CoMeMo can deal with, (b) **CoMeMo Workbench** helps users explore, browse and edit associations in the CoMeMo information bases, (c) **CoMeMo Mobile Termina**l which enables users ubiquitous use of CoMeMo information, (d) **CoMeMo Server** which articulates and serves associations on demand, and mediates between information users and developers, (e) **CoMeMo Protocol** which exchanges information between each component, and (f) **CoMeMo Agent** which executes user tasks, negotiates common problems with other agents, and presents the contents of CoMeMo information bases to users (Figure 2). We have implemented the CoMeMo Incorporator and the CoMeMo Workbench and these will be described as follows.

3.1. CoMeMo Incorporator

The CoMeMo Incorporator gathers and converts a wide variety of information into information representations which CoMeMo can deal with. For example, information is gathered from HTML documents on WWW and newspaper databases [4,5,6]. Figure 3 shows how an algorithm works for HTML documents. It extracts keywords by morphological analysis [7] and generates associations by analyzing HTML structures.

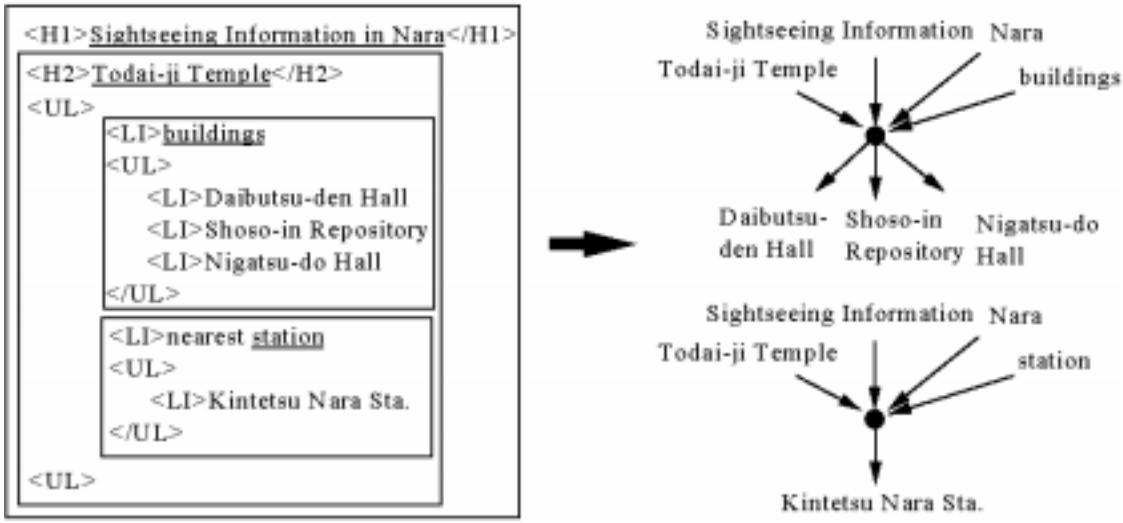

Figure 3. CoMeMo Incorporator

3. 2. CoMeMo Workbench

The CoMeMo Workbench helps users explore, browse and edit associations in the CoMeMo information bases. In addition to basic browsing and editing functions, some facilities which helps users concept formation are stated below.

(a) Information Exploration Facility

Information exploration facility helps users to explore and reorganize information bases.

- **Focus**: to hide units which are unrelated to selected units
- **Neighbor Search**: to display units which are linked to selected units
- **Path Finding**: to display relations between selected units
- **Unit Search**: to display units by keyword search

(b) Information Unification Facility

Information unification facility unifies various associations (e.g. generated from WWW pages or those created by humans) into new associations and removes unnecessary associations using heuristics. These heuristics are based on simple pattern-matching algorithms. Example heuristics are listed below.

- unification of units whose labels are the same
- unification of units referring to user dictionaries
- generation of associations between units when a unit's label is included in another unit's label
- unification of associations whose keys are the same

(c) Information Refinement Facility

Information refinement facility helps users to refine incoherent associations into coherent ones using heuristics.

- **Orthogonal Decomposition**: to decompose a given information base into coherent groups of associations based on intersection of associations
- **Analogical Refinement**: to elaborate information bases based on the measurement of similarity of units.

(d) IS-A Relation Generation Facility

IS-A relation generation facility helps users to generate IS-A generations from given associations by analyzing the class of the units.

(e) Frame Generation Facility

Frame generation facility helps users to generate frames, in other words, a set of entities and attributes from given associations and IS-A relations by path-finding.

4. Test Cases

4.1. Test Case 1: Integration of Creative Thinking and Information Retrieval

Information search is one of important processes in creative thinking. However, most of database systems are separate from creative thinking support systems. We attempt to integrate these two processes: (a) creative thinking and (b) information retrieval.

Information unification facility helps users to utilize information bases possessed by himself/herself or other members in his/her small group. When it is used in bigger communities or societies, ontology and background knowledge bases may be necessary to supply the deficit of the contexts. Figure 4 shows a simple example of how the information unification facility works. The upper left is an workspace written by a user. Associations in information base 1 and 2 are unified and reorganized from the viewpoint of the workspace.

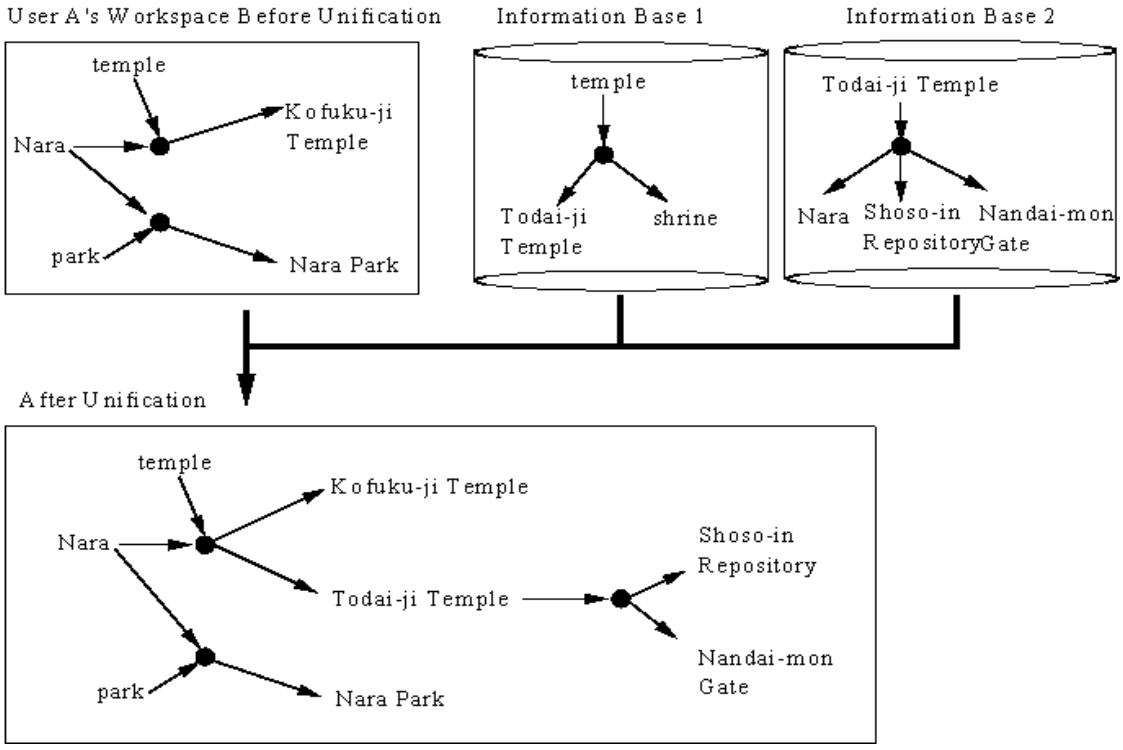

Figure 4. Information Unification

Members of our lab have constructed information bases of slides, survey and so on. User A wrote a plot to write a paper (Figure 5) and then unified information bases in the lab. As a result, a piece of slides or papers contained in User B's survey information base and User A's slide information base are integrated through the workspace (Figure 6).

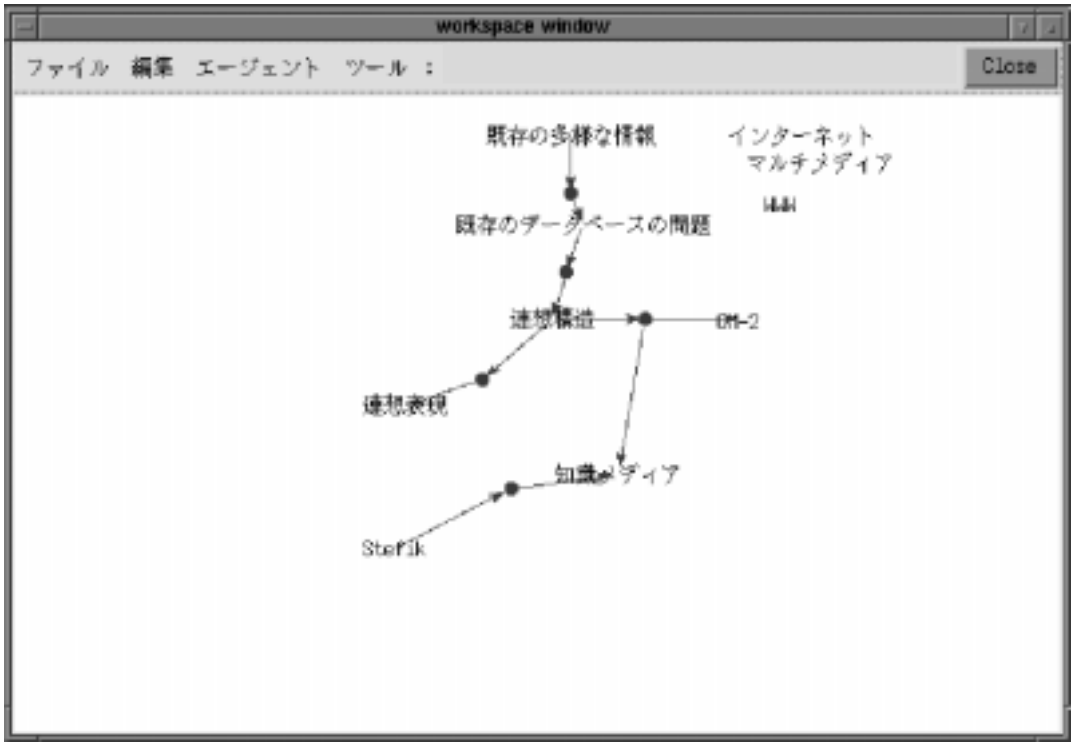

Figure 5. An Example Workspace Before Unification

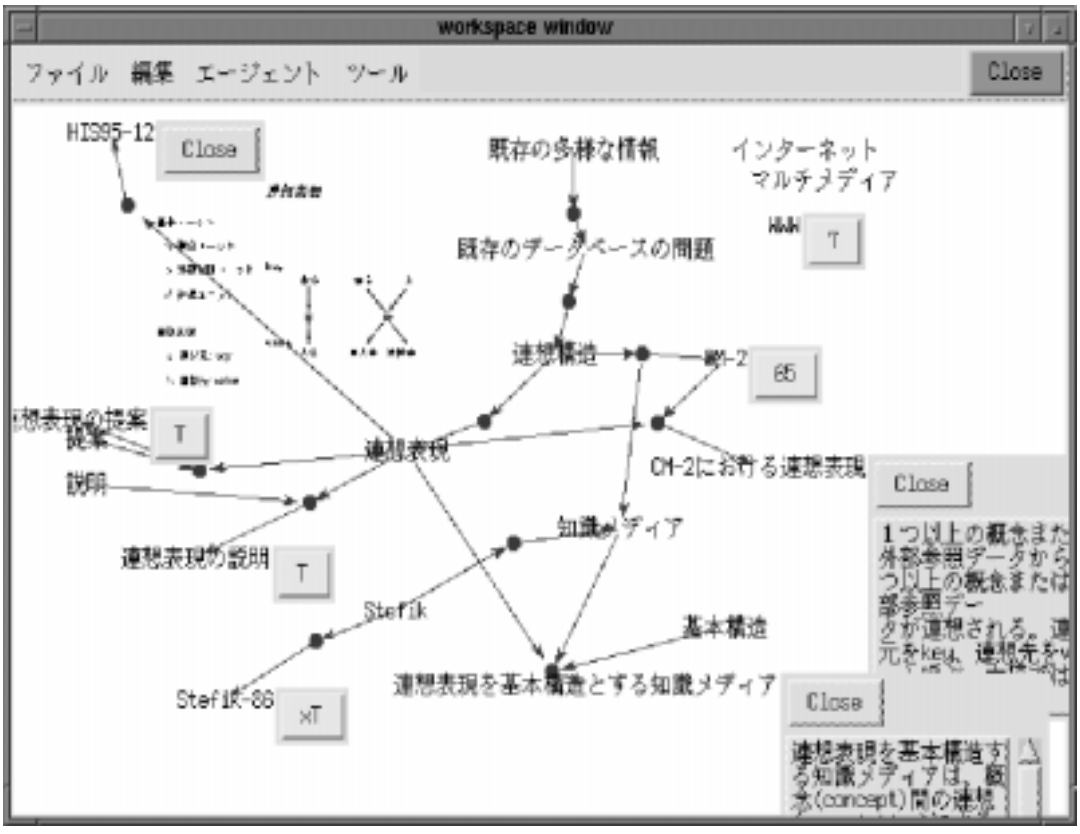

Figure 6. An Example Workspace After Unification

4.2. Test Case 2: Ontology Development from WWW Pages

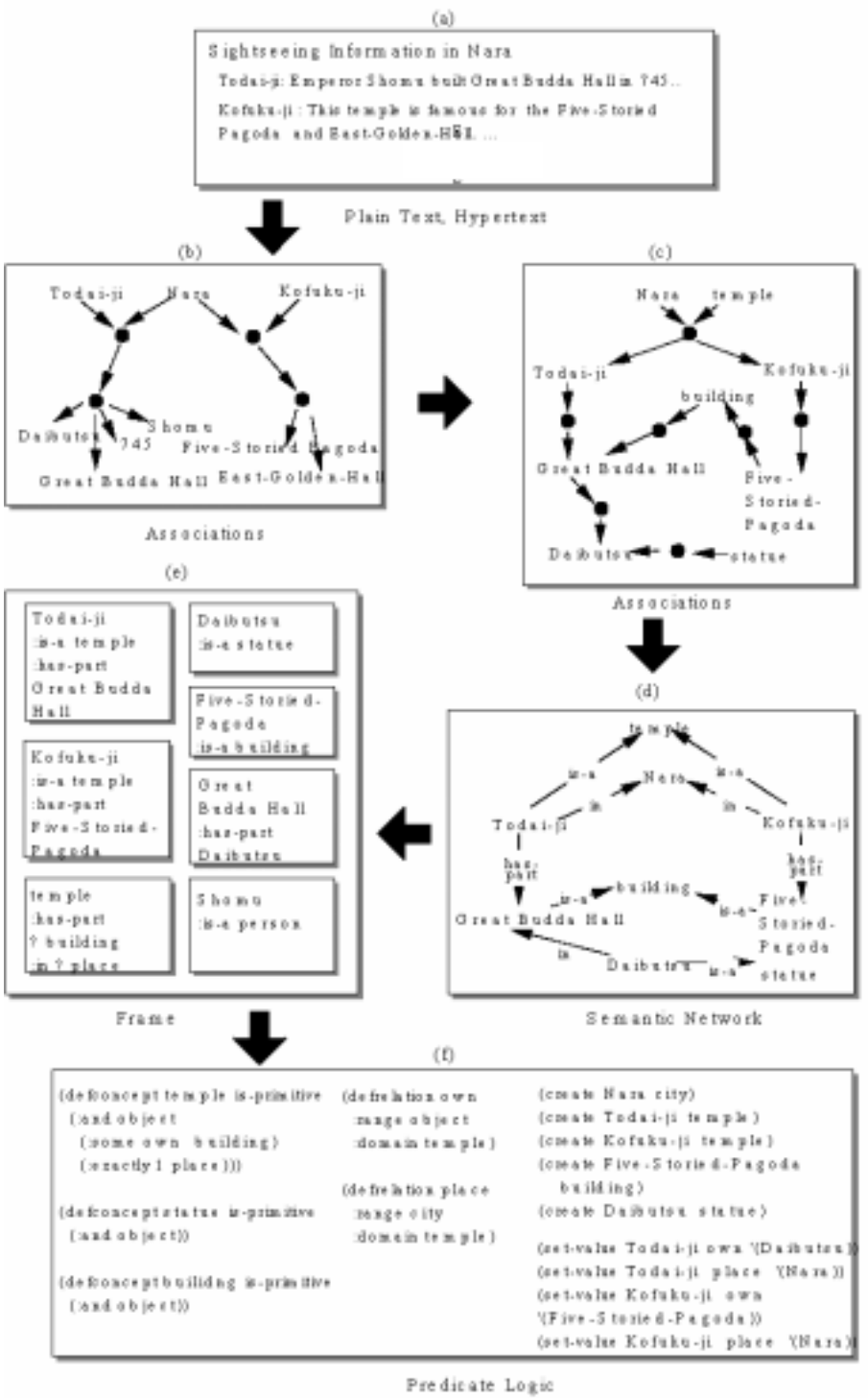

Figure 7. Ontology Development

Ontology development 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 using associations 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 7.

We gave 30 WWW pages concerning ARPA Intelligent Integration of Information (I3) Initiative to the CoMeMo Incorporator. Each page contains overview of projects which belong to the I3 Initiative. The CoMeMo Incorporator generated associations as shown in Figure 8. Generated associations themselves are incoherent and cannot be used as ontology as they are. After associations were unified by Information unification facility, information refinement facility assisted users to form these associations into more coherent structure. Figure 9 shows an example result of refined associations by collaborations between humans and computers on the CoMeMo Workbench concerning the project ``Agent-Based Software Interoperation.'' It corresponds to level c ontology (refined associations) in Figure 7. See [6] for details.

5. Experiment

5.1. Experiment 1: CoMeMo Incorporator

We tested several cases to generate associations from full-text newspaper database, HTML documents and so on. The number of units is around 1000 - 2000. Table 1 shows the results. See [5] for details.

Table 1. Results of Experiment 1

Information Soutces	Test	Precision	Recall
Nikkei Newspaper Full-text Database	Test 1	63 %	91 %
	Test 2	83 %	71 %
HTML Documents on WWW	Test 1	90 %	83 %
	Test 2	68 %	73 %

$$\text{Precision:} \quad \frac{\text{appropriate units}}{\text{generated units}} \times 100$$

$$\text{Recall:} \quad \frac{\text{generated units}}{\text{units which should be extracted}} \times 100$$

5.2. Experiment 2: Information Refinement Facility in CoMeMo Workbench

We manually constructed an information base for regional guide of Nara, Japan. It

contains about 1,850 concepts and 870 associations. Information refinement facility suggested 277 revisions, about 99 of which have been found useful [4].

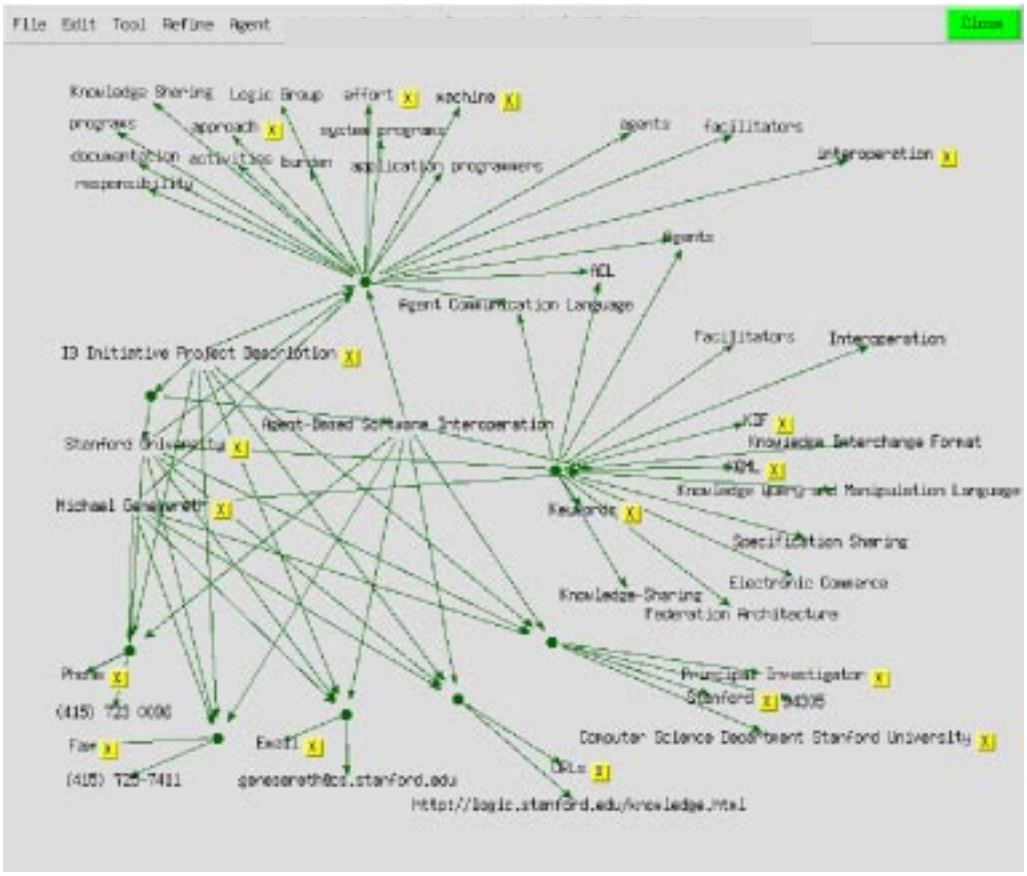

Figure 8. An Example Workspace of Generated Associations

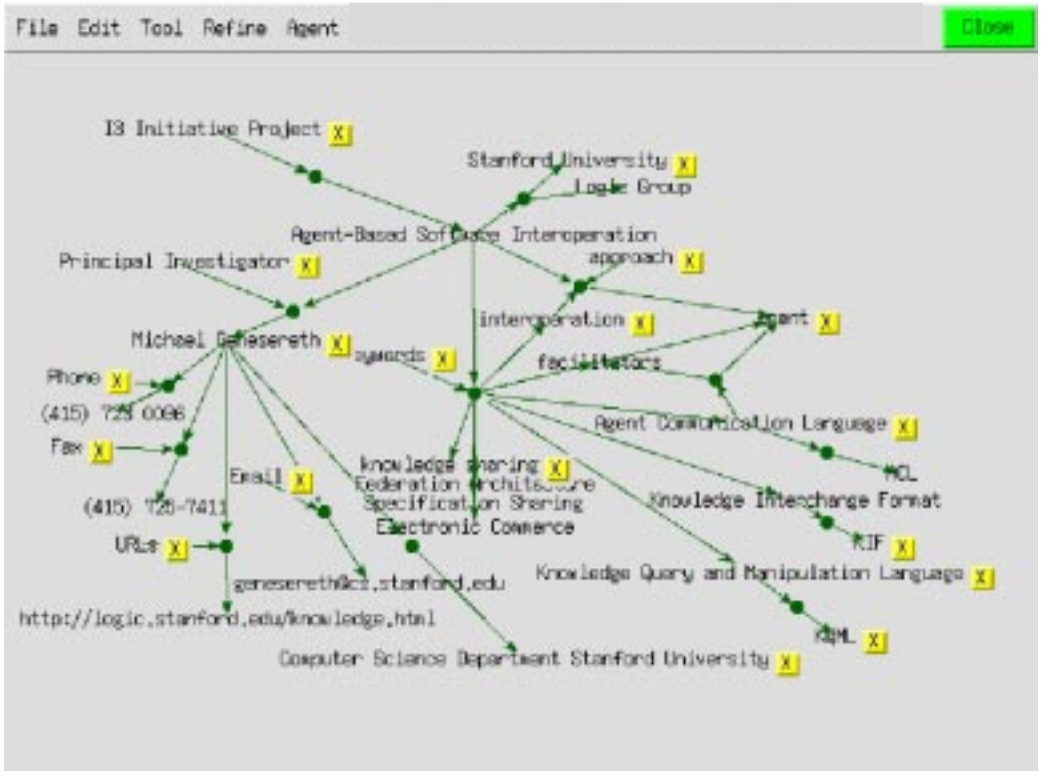

Figure 9. An Example Workspace of Refined Associations

6. Related Work and Discussion

Our work is related to much research work (e.g. [8,9,10,11,12]) on information gathering from the Internet. 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.

Kautz studied the use of agents in assisting and simplifying person-to-person communication for information gathering tasks [13]. They focus on the use of a software agent. We concentrate on the process of how humans create knowledge and information.

Sumi[14] and Hori[15] claim the importance of knowledge and information in the field of creative thinking support.

Huffman and Steiner[16] propose a similar method using heuristic join to combine data from multiple structured sources.

Mizoguchi analyze the problems of existing knowledge bases and propose a method to build task ontologies for knowledge reuse [17].

Our approach is to provide a framework of concept formation with a low structural facilities and to facilitate raw information from vast information sources to be incorporated without much labor and gradually refined and elaborated.

Results of Experiment 1 show that it is easy to generate associations from mark-up documents. However, there is much space to improve as for Experiment 2. Unfortunately, the rate of useful suggestions from the heuristics, being less than 40%, seems to be too low. To improve the quality of heuristics, we are currently looking at introduction of other kinds of heuristics and a domain knowledge. One possibility is introducing a notion of significance of association.

7. Conclusions

We proposed an information representation called associations for concept formation. We described a system called CoMeMo which implements our claim. We reported the evaluation of CoMeMo against two test cases: (1) integration of creating thinking and information retrieval and (2) ontology development from WWW pages.

References

[1] Nonaka, I., and Takeuchi, H. The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation, Oxford University Press, 1995.

[2] Noguchi, Y. Super Information Mangement (CHOSEIRIHOU), Chuko-Shinsho, 1993 (in Japanese).

[3] Freeman, E., and Fertig, S. Lifestreams: Organizing your electronic life. In Working Notes of the 1995 AAAI Falls Symposium on AI Applications in Knowledge Navigation and Retrieval, 1995, 38-44.

[4] Maeda, H., Koujitani, K. and Nishida, T. A Knowledge Media Approach using Associative Representation for Constructing Information Bases. In Proceedings of the Ninth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-96), Fukuoka, Japan, 1996, 117-126.

[5] Maeda, H., Koujitani, K. and Nishida, T. Information Reorganization using Associative Structures. Transactions of Information Processing Society Japan, 38 (3), 1997 (in Japanese).

[6] Koujitani, K. Constructing Ontologies Based on Real Data, Master's Thesis, NAIST, 1996 (in Japanese).

[7] Brill, E. Some Advances in Transformation-based Part of Speech Tagging. In Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), 1994.

[8] Levy, A.Y., Sagiv Y., and Srivasava, D. Towards Efficient Information Gathering Agents. In Working Notes of the 1994 AAAI Spring Symposium on Software Agents, 1995.

[9] Armstrong, R, Freitag, D., Joachims, T., and Mitchell, T. WebWatcher: A Learning Apprentice for the World Wide Web. In Working Notes of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Environments, 1995, 6-12.

[10] Balabanovi'c, M., and Shoham, Y. Learning Information Retrieval Agents: Experiments with Automated Webbrowsing. In Working Notes of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Environments, 1995, 13-18.

[11] Li, W. Knowledge Gathering and Matching in Heterogeneous Databases, In Working Notes of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Environments, 1995, 116-121.

[12] Iwazume, M., Takeda, H., and Nishida, T. Ontology-based Information Gathering and Text Categorization from the Internet. In Proceedings of the Ninth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AID-96), Fukuoka, Japan, 1996, 305-314.

[13] Kauz, H., Selman B., and Milewski, A. Agent Amplified Communication. In Proceedings of the Thirteen National Conference on Artificial Intelligence (AAAI-96), 1996, 3-9.

[14] Sumi, Y., Hori, K., and Ohsuga, S. Computer-aided Thinking Based on Mapping Text-objects into Metric Spaces. In Proceedings of the Second Pacific Rim International Conference on Artificial Intelligence, 1992.

[15] Hori, K, A System for Aiding Creative Concept Formation. IEEE Transactions on Systems, Man and Cybernetics, 24(6), 1994, 882-894.

[16] Huffman S.B., Steier D. Heuristic Joins to Integrate Structured Heterogeneous Data. In Working Notes of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Environments, 1995, 74-77.

[17] Mizoguchi R., Vanwelkenhuysen, J., and Ikeda M. Task Ontology for Reuse of Problem Solving Knowledge. Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing, 1995, 84-94.