A KNOWLEDGE MEDIA APPROACH USING ASSOCIATIVE REPRESENTATION FOR CONSTRUCTING INFORMATION BASES

Harumi Maeda, Kazuto Koujitani and Toyoaki Nishida
Graduate School of Information Science,
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-01 Japan
Email: harumi-m@is.aist-nara.ac.jp

Abstract

In this paper, we present a new approach based on *knowledge medium* using *associative representation* as a framework of information representation to gather raw information from vast information sources and to integrate it into information bases cost-effectively.

We then present a knowledge media information base system called CM-2 which provides users with a means of accumulating, sharing, exploring and refining conceptually diverse information gathered from vast information sources. We describe the system's four major facilities; (a) an *information capture facility*, (b) an *information integration facility*, (d) an *information retrieval facility* and (d) an *information refinement facility*. We discuss the strength and weakness of our approach by analyzing results of experiments.

keywords: associative representation, knowledge media, knowledge media system, CM-2, information base

1 Introduction

There exist various kinds of information sources around us. For instance, personal memoranda, research notes, hypertexts, image files and so on. Most of such information is conceptually diverse in the sense that its semantics is not rigorously defined.

In addition, widespread access to the Internet and WWW has led to a new phase in information acquisition. There already exist large scale information resources and they are increasing rapidly. We need to integrate a wide variety of information into personal information space from our point of view. However, it seems almost impossible to design a well-defined conceptual structure for organizing diverse information obtained from heterogeneous information sources.

In this paper, we present a new approach based on *knowledge medium* [Stefik86] using *associative representation* as a framework of information representation. The basic recognition behind this research is a trade-off between the benefit from conceptually well-structured information space and the cost for organizing information space. The more well-structured information representation becomes, the more useful it is for computational manipulation, however, the more expensive the cost of information acquisition becomes. Associative representation is a plain and weakly structured knowledge medium which is visible and manipulatable to humans and computers. We use associative representation to gather raw information from vast information sources and to integrate it into information bases cost-effectively.

We then present a knowledge media information base system called CM-2 which provides users with a means of accumulating, sharing, exploring and refining conceptually diverse information gathered from vast information sources. We describe the system's four major facilities;

- an *information capture facility* which helps users gather information from multiple information sources
- an *information integration facility* which allows users to integrate heterogeneous information into personal information space from the user's point of view
- an *information retrieval facility* which gives users access to multimedia information stored in the information base through associative indexing mechanisms

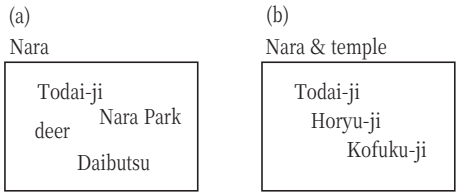

Figure 1: Example associations in CM-2

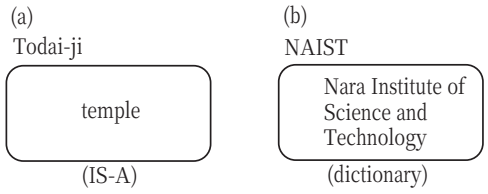

Figure 2: Example associations (special types)

- an *information refinement facility* which helps users reorganize the information base to be more comprehensive

We discuss the strength and weakness of our approach by analyzing results of experiments.

In what follows, we first describe the role of a plain indexing method using associative representation and overview the CM-2 information base system. We then present the system's four major facilities. Finally, we show experimental results and make discussion.

2 Associative Representation in CM-2 Information Base System

2.1 An Indexing Method Using Associative Representation

In this paper, we focus on *associative representation*, which allows the user to explore a way of articulating conceptually diverse information by aggregating conceptually relevant information. The basic entities of associative representation are (a) a *unit* which represents either a concept or an external datum, and (b) an *association* which connects a collection of key concepts (hereafter *keys*) with a collection of units (hereafter *values*) which is normally reminded by the given keys. Figure 1 shows a couple of associations. Figure 1(a) says that given a concept "Nara", one may be reminded of "Todai-ji", "Nara Park", "deer", and "Daibutsu". Figure 1(b) is an example of association with more than one key. It says that "Todai-ji", "Horyu-ji", and "Kofuku-ji" are reminded when "Nara" and "temple" are given as keys.

Users can define special types of associations to be used in information integration and refinement facility. Figure 2(a) is a "IS-A" relation which connects a unit with other units which are reminded as a class of the given unit. Figure 2(b) is a "dictionary" relation which can be used for translation.

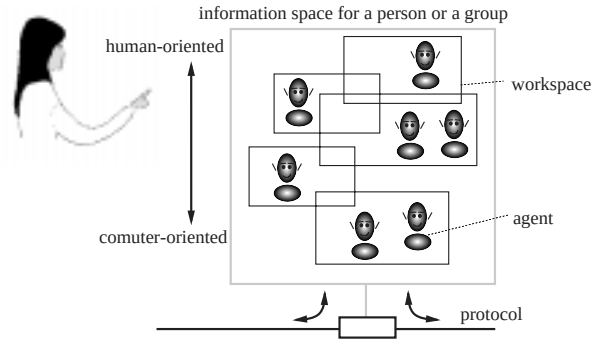

Figure 3: The Architecture of the CM-2 Information Base System

2.2 Overview of the CM-2 Information Base System

CM-2 [1] is a knowledge media information base system which provides users with a means of accumulating, sharing, exploring, and refining conceptually diverse information gathered from vast information sources.

CM-2 consists of a collection of information bases. Each CM-2 information base is possessed by an individual person or a group (Figure 3) and it consists of a collection of *workspaces* and *agents*. Each workspace provides a particular view of multimedia information stored in the information base.

Each agent manipulates information tasks and interacts with the user. The user or the agents in CM-2 can interact with other, or incorporate information from other kinds of information sources connected to the Internet.

Figure 4 shows an example screen of CM-2.

In what follows, we describe four major facilities of CM-2.

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

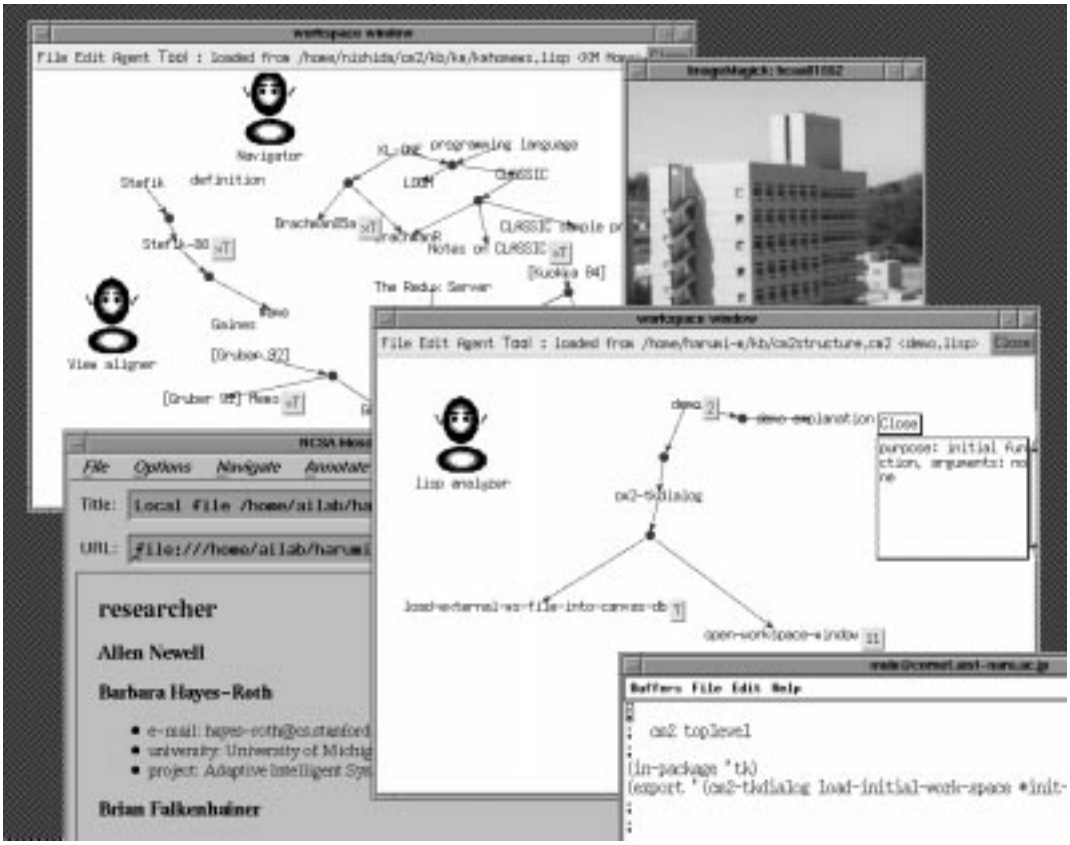

Figure 4: An Example Screen of CM-2

3 Information Capture Facility

The structure of information may well differ according to authors, and people might give different names to same concept and same names to different concepts.

Compare Figure 5(a) and (b). These are WWW pages of two famous AI researchers [2] [3] displayed on WWW browsers. They are different in structure and use of words. We need to obtain useful information from these kinds of diverse information sources.

Information capture facility helps users gather information from multiple information sources and generate CM-2 associations.

It is easy to generate associative representations from various information sources using a simple keyword extraction and text analysis algorithm.

We have implemented capture programs for digitized information, such as UNIX file system, program files written in Lisp, Nikkei newspaper full-text database and HTML documents on WWW.

[2] http://www.cs.rochester.edu/u/james/
[3] http://www-ksl.stanford.edu/people/bhr/index.html

We have also implemented those for capturing undigitized information, such as ideas. Users can input units using keyboard and connect them using mouse through workspaces.

3.1 Information Capture for WWW pages

We focus on capture facility for HTML documents on WWW. The general procedure of the facility is composed of the following steps.

1. generation of raw CM-2 units and associations
   (a) collection of HTML documents by analyzing URL
   (b) extraction of noun phrases and generation of units using Rule Based Tagger by Eric Brill [Brill68]
   (c) generation of associations by analyzing the structure of HTML documents
2. generation of "IS-A" relations and modification of units using domain knowledge

(a) James Allen's Home Page (b) Barbara Hayes-Roth's Home Page

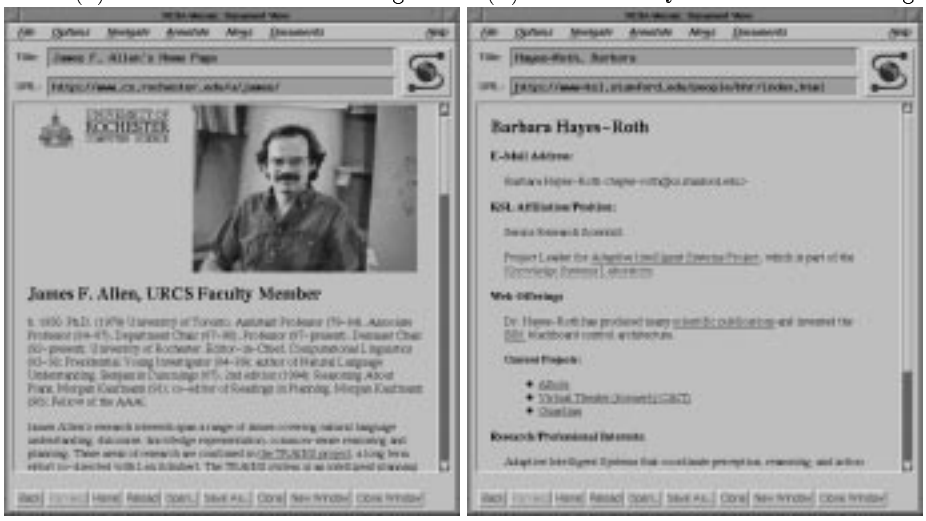

Figure 5: Example Home Pages

(a) extraction and generation of "IS-A" relations from units

(b) removal of unnecessary units and modification of associations

An example of domain knowledge used in step2 (a) is as follows:

> it is inferred that the class of a unit "James F. Allen" is a "people", because the unit's label contains "James" which is one of common English names.

Figure 6 shows the overview of the algorithm when a URL of James Allen's Home Page is given.

4 Information Integration Facility

Information integration facility allows users to integrate heterogeneous information into personal information space from the user's point of view. When a user input items for reorganizing information, it generates new associations in accordance with users request and itemizes them.

The following describes the general procedure of the facility.

1. unification of units and associations using several heuristics

2. generation of new associations according to user input

   (a) extraction of keys by path finding

   (b) extraction of values by path finding

   (c) generation of new associations

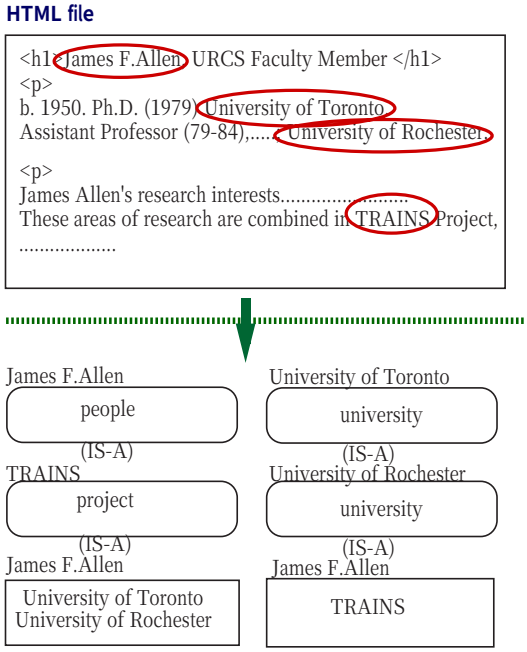

Figure 6: Information Capture Facility

Some examples of heuristics used in step1 are stated 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

Let us think a case that when a user wants to know AI researchers and their contact information and projects concerning your research interest but there is no such database available. The user may search WWW pages about AI, read appropriate encyclopedia and organize the information using his/her knowledge. Information integration facility helps user's such process.

Figure 7 shows how the facility answer the following question against the sets of CM-2 associations which are mixtures of associations generated by information capture facility and those obtained by other information sources.

> "Display a list of researchers and related projects concerning 'reasoning' ?"

Figure 8 and figure 9 illustrate example results of the facility.

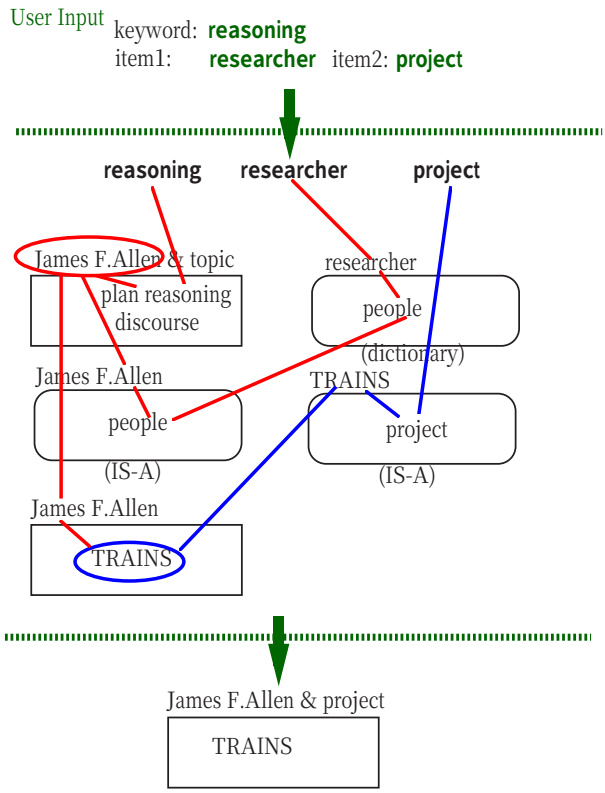

Figure 7: Information Integration Facility (step2)

5 Information Retrieval Facility

Information retrieval facility gives users access to multimedia information stored in the information base through associative indexing mechanisms. The system has three information retrieval facilities: (a) keyword search, (b) neighbor search and (c) intelligent associative retrieval. The rest part of this section describes neighbor search and intelligent associative retrieval.

5.1 Neighbor Search

Neighbor search enables users to search and display units which are linked to the selected unit by associations. For example, when an association shown in Figure 1 is given and the user selects "Nara", linked units such as "temple", "Todai-ji", "Horyu-ji" and "Kofuku-ji" will be displayed. Users can execute neighbor search by pressing buttons displayed nearby units on workspaces [4].

Neighbor Search causes a problem when there are too many values associated with the selected unit; it is very difficult to identify the displayed units. Figure 10 shows an example of workspace in such a case. To remedy this problem, we need more intelligent and dynamic search facility to obtain the desired information and it will be described in the next section.

5.2 Intelligent Associative Retrieval

Path finding is a powerful means of retrieving information, in particular when what is contained in an information base is structurally different from the presupposition of a given query.

Figure 11 illustrates how the algorithm works to answer a question:

> "are there any places in Nara that are famous for rhododendron?"

[4] These buttons are displayed when units have some values undisplayed on workspaces. A number displayed within buttons describes the number of values of the unit.

Figure 8: An Example Screen of Information Integration Facility 1 (A List of AI Researchers with their Related Information)

Figure 9: An Example Screen of Information Integration Facility 2 (A List of AI Projects with their Related Information)

Intelligent associative retrieval is based on the idea of "spreading activation" on semantic networks [Quillian68] and composed of the steps below.

1. extraction of units by analyzing the query
2. extraction of answers by path finding
3. display of the extracted units, answers, paths which links these concepts, and concepts which are useful for understanding the paths

Figure 12 shows the result of the intelligent associative retrieval.

6 Information Refinement Facility

Information refinement facility helps users reorganize the information base to be more comprehensive.

Compare two sets of associations in Figure 1 (a) and (b). The association in Figure 1(b) is more comprehensive and useful than that in Figure 1(a) because various kinds of entities are mixed up in the association in Figure 1(a).

We present a couple of heuristic techniques which will detect inappropriate associations from CM-2 information base and suggest a possible way of remedying them.

Figure 10: An Example Screen of Neighbor Search (Problematic Case)

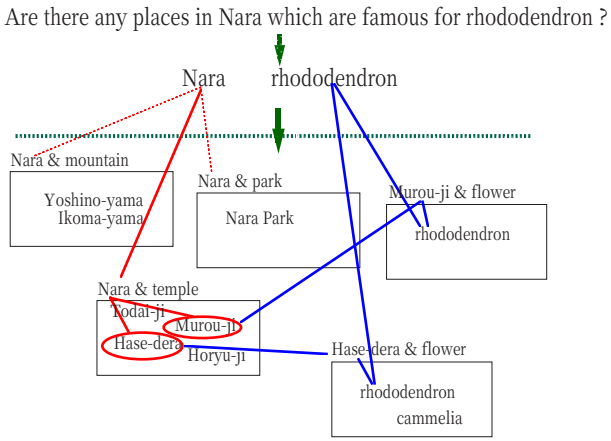

Figure 11: Intelligent Associative Retrieval

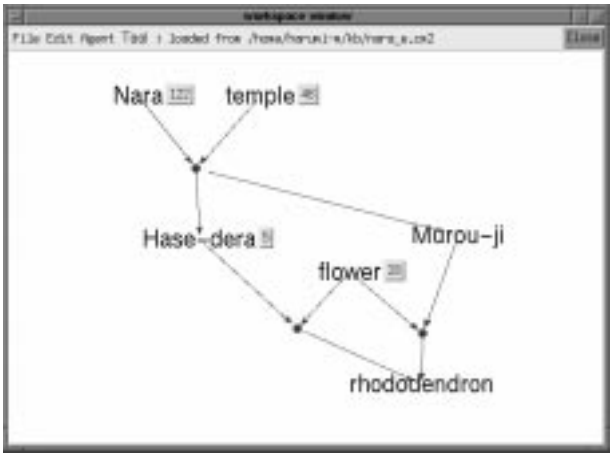

Figure 12: An Example Screen of Intelligent Associative Retrieval

6.1 Orthogonal Decomposition

Orthogonal decomposition attempts to decompose a given information base into coherent groups of associations, by analyzing how the user intersects associations. It is a technique of refining CM-2 information base using diagnosis rules shown in Figure 13.

An example of orthogonal decomposition is illustrated in Figure 14.

6.2 Analogical Refinement

Analogical refinement is a less efficient but more powerful technique for further elaborating information base based on the measurement of similarity.

Given a couple of non-orthogonal keys x and y, we

if
$\langle$ for concepts x and y:
$\mathrm{V}^*[\{x\}] \cap \mathrm{V}^*[\{y\}] \neq \mathrm{V}^*[\{x,y\}]$ $\rangle$
then
penalty $\leftarrow \frac{|(\mathrm{V}^*[\{x\}] \cap \mathrm{V}^*[\{y\}]) - \mathrm{V}^*[\{x,y\}]|}{|\mathrm{V}^*[\{x,y\}]|}$;
suggestion $\leftarrow$ "resolve the difference between $\mathrm{V}^*[\{x\}] \cap \mathrm{V}^*[\{y\}]$ and $\mathrm{V}^*[\{x,y\}]$, by adding z to $\mathrm{V}[\{x,y\}]$ if $z \notin \mathrm{V}^*[\{x,y\}]$ and $z \in (\mathrm{V}^*[\{x\}] \cap \mathrm{V}^*[\{y\}])$ "

if
$\langle$ for two sets of concepts α, β, $\alpha \subset \beta$:
$\exists z[\, z \in \mathrm{V}[\alpha] \wedge z \in \mathrm{V}[\beta]\,]$ $\rangle$
then
penalty $\leftarrow \infty$;
suggestion $\leftarrow$ "remove z from $\mathrm{V}[\alpha]$."

Figure 13: Diagnosis rules for orthogonal decomposition

define the similarity $\mathrm{Sim}[x,y]$ as shown in Figure 15. Based on that definition, we define the key similarity $\mathrm{Sim}^*[\alpha,\beta]$ between keys α and β as the sum of maximal pairwise similarities of units in α and β. Namely,

$$\mathrm{Sim}^*[\alpha,\beta] = \max\left[\sum_{x\in\alpha}\max_{y\in\beta}[\mathrm{Sim}[x,y]], \sum_{y\in\beta}\max_{x\in\alpha}[\mathrm{Sim}[x,y]]\right]$$

For concepts x, y, and a threshold $\theta > 0$, we denote $x \sim y$ if $\mathrm{Sim}[x,y] \geq \theta$. Similarly, for keys α, β, and a threshold θ, $\alpha \sim \beta$ if $\mathrm{Sim}^*[\alpha,\beta] \geq \theta$.

The analogical refinement heuristic suggests to refine a CM-2 information base according to the following diagnosis rule:

if
$x \in \mathrm{V}^*[\alpha]$,
$y \in \mathrm{V}^*[\beta \cup \{a\}]$, and
$x \notin \mathrm{V}^*[\alpha \cup \{a\}]$
then
penalty $\leftarrow \mathrm{Sim}[x,y] + \mathrm{Sim}^*[\alpha,\beta]$
suggestion $\leftarrow$ add x to $\mathrm{V}[\alpha \cup \{a\}]$.

There are several interesting suggestions. For example, from "cherry blossom" $\in$ V[{"Ikoma park"}] and, "cherry blossom" $\in$ V[{"flowers"}], we obtained

"cherry blossom" $\in$ V[{"Ikoma park", "flowers"}],

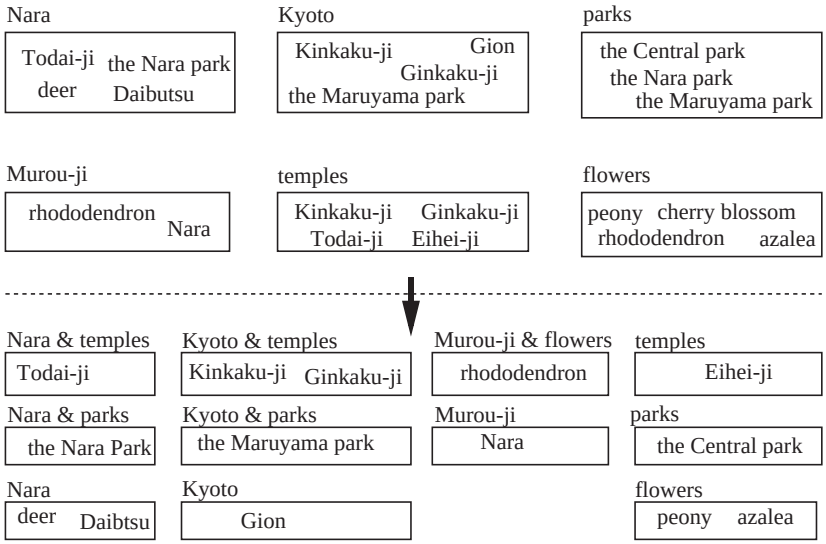

Figure 14: Orthogonal decomposition of CM-2 information base

Given a couple of non-orthogonal keys x and y, we define the similarity $\mathrm{Sim}[x,y]$ between x and y from three perspectives and let it:

$$\mathrm{Sim}[x,y] = \frac{\mathrm{Sim}^{(\mathrm{a})}[x,y] + \mathrm{Sim}^{(\mathrm{b})}[x,y] + \mathrm{Sim}^{(\mathrm{c})}[x,y]}{3} \in [0,1].$$

$\mathrm{Sim}^{(\mathrm{a})}[x,y]$ measures the similarity between x and y by comparing concepts in $\mathrm{V}^*[\{x\}]$ and those in $\mathrm{V}^*[\{y\}]$. The definition is as follows:

$$\begin{aligned}\mathrm{Sim}^{(\mathrm{a})}[x,y] = \ & \frac{1}{|\mathrm{V}^*[\{x\}] \cup \mathrm{V}^*[\{y\}]|} \\ & \cdot\Big(\big|\{z \mid z \in \mathrm{V}^*[\{x\}] \wedge z \in \mathrm{V}^*[\{y\}]\}\big| \\ & \quad + \big|\{z \mid z \in \mathrm{V}^*[\{x\}] - \mathrm{V}^*[\{y\}] \wedge \exists u[\, u \in \mathrm{V}^*[\{y\}] \wedge (\mathrm{K}^*[z] \cap \mathrm{K}^*[u] \neq \{\})\,]\}\big| \\ & \quad + \big|\{z \mid z \in \mathrm{V}^*[\{y\}] - \mathrm{V}^*[\{x\}] \wedge \exists u[\, u \in \mathrm{V}^*[\{x\}] \wedge (\mathrm{K}^*[z] \cap \mathrm{K}^*[u] \neq \{\})\,]\}\big| \Big).\end{aligned}$$

$\mathrm{Sim}^{(\mathrm{b})}[x,y]$ measures the rate of common keys of associations containing x and y as values. Namely,

$$\mathrm{Sim}^{(\mathrm{b})}[x,y] = \frac{|\{z \mid z \in \mathrm{K}^*[x] \wedge z \in \mathrm{K}^*[y]\}|}{|\mathrm{K}^*[x] \cup \mathrm{K}^*[y]|}.$$

$\mathrm{Sim}^{(\mathrm{c})}[x,y]$ measures the rate of keys orthogonal both to x and to y. Thus,

$$\mathrm{Sim}^{(\mathrm{c})}[x,y] = \frac{|\{z \mid \langle z \text{ is orthogonal to } x\rangle \wedge \langle z \text{ is orthogonal to } y\rangle\}|}{|\{z \mid \langle z \text{ is orthogonal to } x\rangle\} \cup \{z \mid \langle z \text{ is orthogonal to } y\rangle\}|}.$$

Figure 15: Defining similarity between concepts

from which we in turn obtained

“iris” $\in$ V[{“Ayameike park”, “flowers”}]

based on

“iris” $\in$ V[{“Ayameike park”}],

“Ikoma park” $\sim$ “Ayameike park”, and

“cherry blossom” $\sim$ “iris”,

as shown in Figure 16.

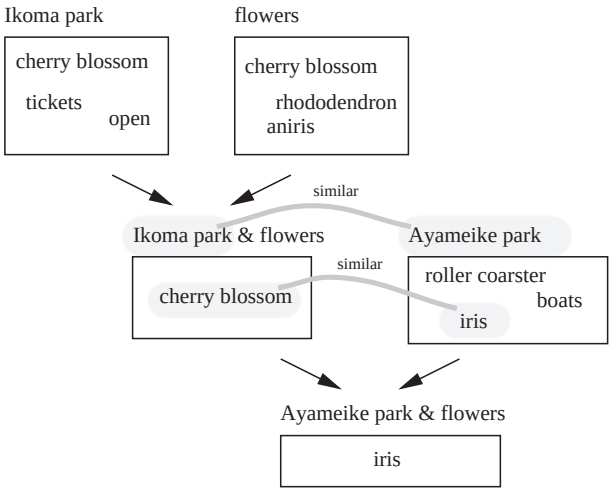

Figure 16: An interesting suggestion obtained by orthogonal decomposition and analogical refinement

7 Experiments

We have implemented CM-2 on top of Common Lisp and tcl/tk. We are evaluating CM-2 against accumulating various kinds of information such as research memoranda, technical surveys, regional guide, personal diary, and so on. Besides testing against these small examples and the examples described so far, we have made a couple of experiments with a nontrivial scale.

Experiment 1: Information Capture Facility We have gathered 22 WWW pages concerning AI researchers. CM-2 has extracted units about 7 classes such as researchers, topics and universities and generated associative representations. We have evaluated that 222 units out of 272 are appropriate.

Experiment 2: Information Integration Facility After having modified CM-2 associations generated in Experiment 1, we have tested information integration facility. A sample question is: "Display AI projects with their related researchers and universities." 60 units out of 72 are extracted properly.

Experiment 3: Information Retrieval Facility We have manually constructed a CM-2 information base for regional guide of Nara, Japan. It contains about 1,850 units and 870 associations. We have tested intelligent associative retrieval against the above information base. We have obtained appropriate 24 answers out of 30 queries.

Experiment 4: Information Refinement Facility We have tested orthogonal decomposition and analogical refinement against the information base constructed in Experiment 3. As a result of orthogonal decomposition, CM-2 has produced 212 revisions, about 80 of which have been found useful. Others are uninteresting. On the other hand, the analogical refinement heuristic has generated 65 suggestions, 20 of which are found useful.

8 Related Work and Discussion

The work reported in this paper is part of the **Knowledgeable Community** [Nishida93] project which aims to develop a computational framework of collecting, accumulating, systematizing, sharing, and creating knowledge by human-computer interaction. Crucial issues in the Knowledgeable Community are (a) knowledge media, (b) ontology and (c) agent-assisted mediation technology. We focus on knowledge media and have built an information base system using associative representation.

Our work is related to recent work on information gathering from heterogeneous sources on Internet ([Levy94],[Armstrong95],[Balabanovi'c95],[Li95]). 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.

The basic recognition behind this research is a tradeoff between the benefit from conceptually well-structured information space and the cost for organizing information space. The more well-structured information representation becomes, the more useful it is for computational manipulation, however, the more expensive the cost of information acquisition and integration becomes.

Our approach is to provide a framework of information representation 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 as more insights are obtained.

How successful is our approach? Experiment 1,2 and 3 have ended up in very promising results. Members of our group have been able to use associative representation to accumulate and access varieties of information taken from vast information sources and access relevant information.

However, Experiment 4 shows that there is much

space to improve heuristics about information refinement facility, since the rate of useful suggestions from the heuristics seems to be low. To improve the quality of heuristics, we are currently looking at introduction of other kinds of heuristics and domain knowledge.

9 Conclusions

In this paper, we have proposed a new approach based on *knowledge medium* using *associative representation* as a framework of information representation to facilitate raw information from vast information sources to be incorporated without much labor.

We have presented CM-2 information base system which provides users with a means of accumulating, sharing, exploring and refining conceptually diverse information gathered from vast information sources. We have described the system's four major facilities: (a) an *information capture facility* which helps users gather information from multiple information sources, (b) an *information integration facility* which allows users integrate heterogeneous information into personal information space from the user's point of view, (c) an *information retrieval facility* which gives users access to multimedia information stored in the information base through associative indexing mechanisms, and (d)an *information refinement facility* which helps users reorganize the information space to be more comprehensive. We have discussed the strength and weakness of the method on the analysis of experimental results.

We have implemented a kernel of CM-2 on top of Common Lisp and tcl/tk. The system currently operates on the UNIX platform.

References

[Armstrong95] Robert Armstrong and Dayne Freitag and Thorsten Joachims and Tom Mitchell. 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.

[Balabanovi'c95] 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.

[Brill68] Eric Brill, Some Advance in Transformation-Based Part of Speech Tagging. In *Proceedings of the Twelefth National Conference on Artificial Intelligence (AAAI-94)*,1994.

[Levy94] Alon Y. Levy and 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.

[Li95] 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.

[Nishida93] Toyoaki Nishida and Hideaki Takeda. Towards the knowledgeable community. In *Proceedings of International Conference on Building and Sharing of Very Large-Scale Knowledge bases 93*,pages 157–166. Japan Information Processing Development Center, 1993.

[Quillian68] M.R.Quillian, Semantic memory. In Marvin Minsky edition, *Semantic Information Processing*, MIT Press, 1968.

[Stefik86] Mark Stefik, The next knowledge medium, *AI Magazine*, 7(1):34–46,1986.