??, ??, 1–13 (??)

# Constructing Information Bases using Associative Structures

**HARUMI MAEDA**
*harumi@media.osaka-cu.ac.jp*
*Media Center, Osaka City University*
*3-3-138, Sugimoto, Sumiyoshi, Osaka 558-8585 JAPAN*

**KAZUTO KOUJITANI**
*kazuto-k@zoo.ncl.omron.co.jp*
*Fuzzy Technology and Business Promotion Division, OMRON Corporation*
*Shimokaiinnji, Nagaokakyo-City, Kyoto, 617-8510 JAPAN*

**TOYOAKI NISHIDA**
*nishida@is.aist-nara.ac.jp*
*Graduate School of Information Science, Nara Institute of Science and Technology*
*8916-5 Takayama, Ikoma, Nara, 630-0101 Japan*

Received ??. Revised ??.

**Abstract.**

We present an approach based on *knowledge medium* using *associative structures* as a framework of information representation to gather raw information from heterogeneous 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 structures, knowledge media, 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 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.

We present an approach based on *knowledge medium* [1] using *associative structures* 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 structures connect a wide variety of information media such as natural language texts, hypertexts and im-

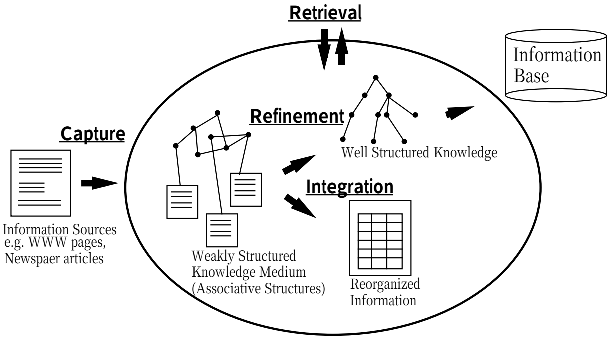


*Fig. 1.* Overview of the Approach

ages without defining the semantics to integrate heterogeneous information.

We use associative structures to gather raw information from vast information sources and to integrate it into information bases cost-effectively. Fig.1 shows the overview of our approach.

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 reorganize heterogeneous information from the user's point of view
- an *information retrieval facility* which gives users access to the information base through associative indexing mechanisms
- an *information refinement facility* which helps users to refine incoherent associations into coherent ones

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

In what follows, we first describe associative structures and give an overview of the CM-2 information base system. We then present the system's four major facilities. Finally, we show experimental results and present discussion.

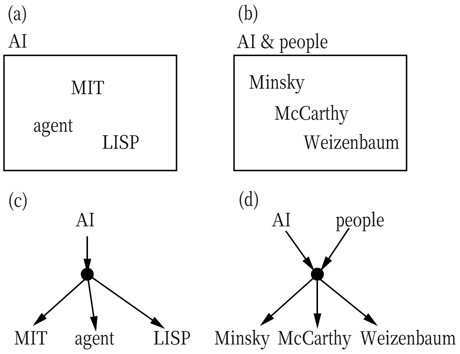


*Fig. 2.* Example associations

## 2. Associative Structures and CM-2 Information Base System

### 2.1. Associative Structures

*Associative structures* allow the user to explore a way of articulating conceptually diverse information by aggregating conceptually relevant information. The basic entities of associative structures are a *unit* which represents either a concept (a label to the universe of discourse) or an external datum (a pointer to multimedia files), and an *association* which connects a collection of key concepts (hereafter *keys*) with a collection of units (hereafter *values*) which are normally reminded by the given keys.

Fig.2(a)(b) shows a couple of associations. Fig.2(a) says that given a concept “AI”, one may be reminded of “MIT”, “agent”, and “LISP.” Fig.2(b) is an example of association with more than one key. It says that “Minsky”, “McCarthy”, and “Weizenbaum” are reminded when “AI” and “people” are given as keys.

Fig.2(c)(d) shows different expression of associations called “dot description” in which a dot describes an association and associated arrows represent direction of the association.

Users can define two special types of associations. (a) IS-A relations which connect a unit with other units which are reminded as a class of the given unit. (b) Dictionary relations describe synonyms which can be used for translation.

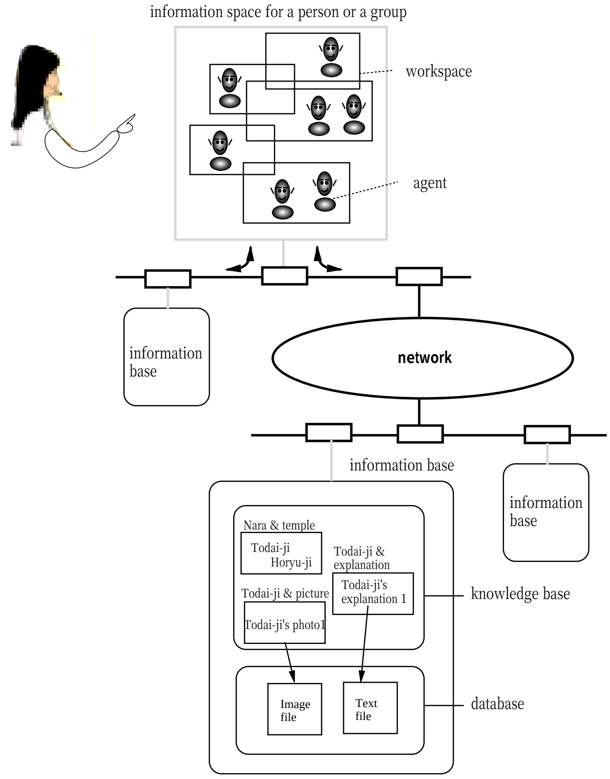


*Fig. 3.* The Architecture of the CM-2 Information Base System

### *2.2.* 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. Fig. 3 shows the architecture of the system.

CM-2 consists of a collection of information bases. Each information base is possessed by an individual person or a group 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 agent can interact with one another, or incorporate information from other kinds of information sources connected to the Internet. Fig.5 shows an example screen of CM-2.

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

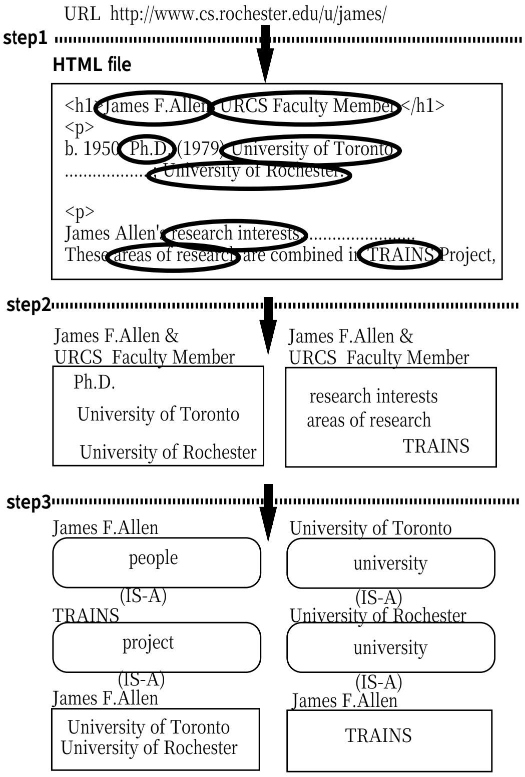


*Fig. 4.* Information Capture Facility

## 3. Information Capture Facility

Information capture facility helps users gather information from multiple information sources and generate associative structures.

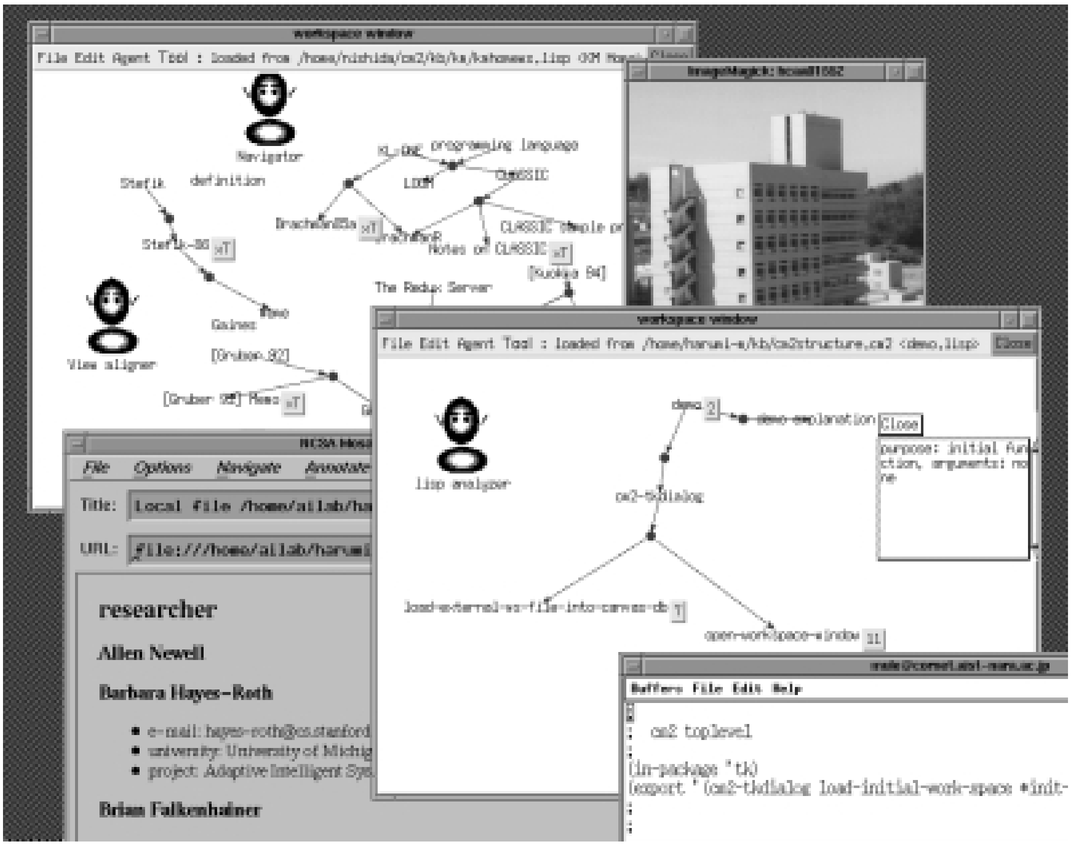
It is easy to generate associative structures from various information sources using a simple keyword extraction and text analysis algorithm.

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 the WWW.

### *3.1. Information Capture from WWW pages*

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

- **Step 1** collection of HTML documents
- **Step 2** generation of units and associations
  - **Step 2.1** generation of units using morphological analysis and heuristics



Title: Local file /home/ailab/ha

URL: file:///home/ailab/harumi

researcher

Allen Newell

Barbara Hayes-Roth

- e-mail: hayes-roth@cs.stanford
- university: University of Michig
- project: Adaptive Intelligent Sys

Brian Falkenhainer

purpose: initial function, arguments: none

```
;
;   cm2 toplevel
;
(in-package 'tk)
(export '(cm2-tkdialog load-initial-work-space *init-
;
;
```

*Fig. 5.* An Example Screen of CM-2

  - **Step 2.2** generation of associations by analyzing HTML structures
- **Step 3** generation of IS-A relations
  - **Step 3.1** generation of IS-A relations using heuristics
  - **Step 3.2** removal of unnecessary units and modification of associations (optional)

**Step 1: Collection of HTML Documents** The system collects HTML documents according to user's input. Users can define one URL to start collection and number of WWW pages which are linked by the first page and select one language (either English or Japanese).

**Step 2: Generation of Units and Associations** Collected HTML documents are analyzed morphologically (by Brill's Rule Based Tagger[2] for English, by JUMAN for Japanese), a series of nouns are extracted and concepts are generated.

After concepts are generated, associations are generated by analyzing HTML tags.

**Step 3: Generation of IS-A relations** IS-A relations are generated from units using several heuristics which identifies the class of a given concept. Some examples are stated below.

- Identify-People Heuristic identifies that a given unit's class is people if the label of the unit is included in a given set of human names.
- Identify-Project Heuristic identifies that a given unit's class is project if (1) a label of the unit (hereafter *label*) contains "project" or (2) all characters included in the label are capital and the label is comprised of more than 3 characters, and, the label is not included in a given set of unnecessary words.

**Step 3.2** is an optional step which aims to reduce the number of associative structures in an information base. Units and associations which are not referred by IS-A relations are treated unnecessary and removed from the information base.

```
begin
  key1-candidates :=
  units linked from item1 by IS-A relations;
  if keyword is not null
    then
      define key1s out of key1-candidates
      by path-finding units whose labels
      include keyword
    else key1s := key1-candidates;
  for any key1 := key1s do
    for any item := items except for
    the first one
      do begin
        key2 := item;
        value-candidates :=
        units linked from key2 by
        IS-A relations;
        define values out of value-candidates
        by path-finding units linked by key1;
        generate an association whose key are
        key1 and key2
        end;
  display the result
  end.
```

*Fig. 6.* Algorithm of Information Integration Facility (Step 2)

**Example** Fig.4 shows how the system works when the URL of James Allen's Home Page [2] is given.

After collecting an HTML file according to the URL, nouns and noun phrases are extracted from it, and concepts are generated (e.g.“James F.Allen”,“URCS Faculty Member”, “Ph.D”,“University of Toronto”and“University of Rochester”) using morphological analysis.

An association is generated whose keys are “James F.Allen”and “URCS Faculty Member”and values are“Ph.D”,“University of Toronto”and“University of Rochester” by analyzing HTML structure.

“James F.Allen” is inferred as “people” by identifying-people heuristic since it includes “James” which is a common English name, and an “IS-A” relation is generated.

Likewise, “University of Toronto” and “University of Rochester” are inferred as “university” by identifying-university heuristic since they include “university” and “IS-A relations” are generated.

“URCS Faculty Member” and “Ph.D” are removed from the information base and from the above association whose keys are “James F.Allen” and “URCS Faculty Member” because they are not referred by IS-A relations.

## 4. Information Integration Facility

Information integration facility allows users to reorganize heterogeneous information from the user's point of view. When a user input items for reorganizing information, it generates new associations in accordance with users input and displays the result in various formats such as lists, tables and networks.

### *4.1. General Procedure*

The following describes the general procedure of the facility.

- **Step 1** unification of units and associations using heuristics
- **Step 2** generation of new associations using path-finding and display of the result

### *4.2. Unification of Units and Associations using Heuristics*

In **Step 1**, units and associations which are generated separately in information capture facility are unified using heuristics.

Some examples of heuristics 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

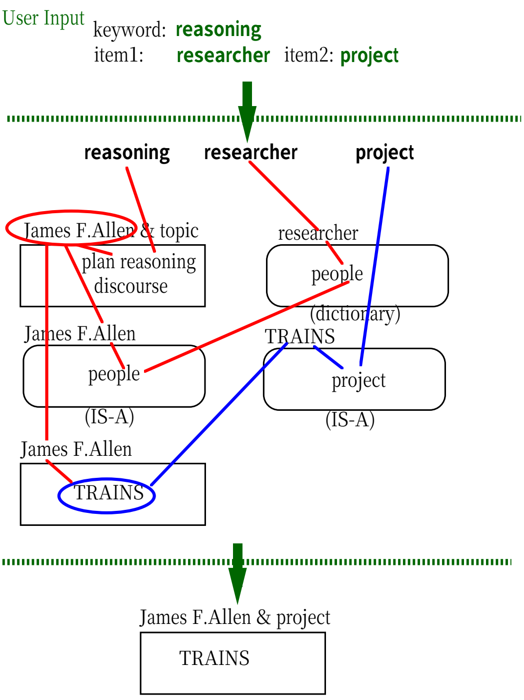
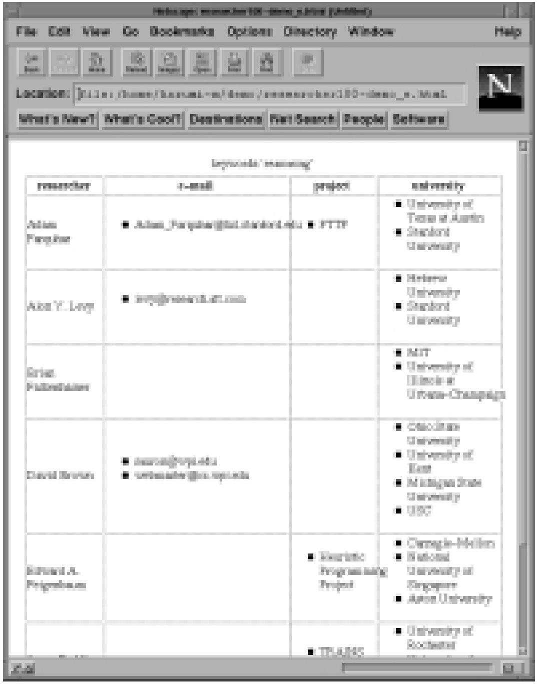


*Fig. 7.* Information Integration Facility (Step 2)

keyword for extraction: "reasoning"
items: "researcher" "e-mail" "project" "university"


Netscape: researcher100-demo_e.html (Untitled)
File Edit View Go Bookmarks Options Directory Window Help
Location: file:/home/harumi-m/demo/researcher100-demo_e.html
What's New? What's Cool? Destinations Net Search People Software

keywords 'reasoning'

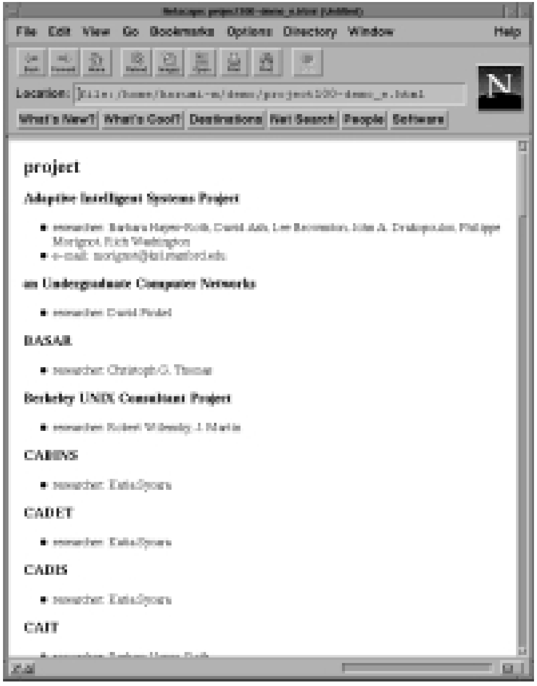
| researcher | e-mail | project | university |
|---|---|---|---|
| Adam Farquhar | ■ Adam_Farquhar@ksl.stanford.edu | ■ FTTP | ■ University of Texas at Austin<br>■ Stanford University |
| Alon Y. Levy | ■ levy@research.att.com | | ■ Hebrew University<br>■ Stanford University |
| Brian Falkenhainer | | | ■ MIT<br>■ University of Illinois at Urbana-Champaign |
| David Brown | ■ dcbrown@wpi.edu<br>■ webmaster@cs.wpi.edu | | ■ Ohio State University<br>■ University of Kent<br>■ Michigan State University<br>■ USC |
| Edward A. Feigenbaum | | ■ Heuristic Programming Project | ■ Carnegie-Mellon<br>■ National University of Singapore<br>■ Aston University |
| | | ■ TRAINS | ■ University of Rochester |

*Fig. 8.* An Example Result of Information Integration Facility (A Table of AI Researchers)

### *4.3. Generation of New Associations using Path-Finding and Display of the Result*

The algorithm of **Step 2** is illustrated in Fig. 7.

keyword for extraction: none
items: "project" "researcher" "e-mail" "university"


Netscape: project100-demo_e.html (Untitled)
File Edit View Go Bookmarks Options Directory Window Help
Location: file:/home/harumi-m/demo/project100-demo_e.html
What's New? What's Cool? Destinations Net Search People Software

project

**Adaptive Intelligent Systems Project**

- researcher: Barbara Hayes-Roth, David Ash, Lee Brownston, John A. Drakopoulos, Philippe Morignot, Rich Washington
- e-mail: morignot@ksl.stanford.edu

**an Undergraduate Computer Networks**

- researcher: David Finkel

**BASAR**

- researcher: Christoph G. Thomas

**Berkeley UNIX Consultant Project**

- researcher: Robert Wilensky, J. Martin

**CABINS**

- researcher: Katia Sycara

**CADET**

- researcher: Katia Sycara

**CADIS**

- researcher: Katia Sycara

**CAIT**

- researcher: Barbara Hayes-Roth

*Fig. 9.* An Example Result of Information Integration Facility (A List of AI Projects)

*Fig. 10.* An Example Screen of Neighbor Search (Problematic)

First item input by users defines user's point of view, or object class. Other items define attribute classes of the object class. Units which are instances of the object class and those which are instances of the attribute classes.

When a keyword is given by the user, only units whose labels contain string of the keyword are extracted among instances of the object class, and related attribute information is reorganized.

```
begin
  analyze query to extract concepts;
  n := 1;
  while (n < 3) and (answers are null) do
    begin
      add units within distance n from the first
        concept to answers;
      for any concept2 := concepts except for
        the first do
          answers := intersection of answers
          and units with distance n
          from concept2;
      n := n + 1
    end;
  display answers, concepts and paths
end.
```

*Fig. 11.* Algorithm of Information Retrieval Facility

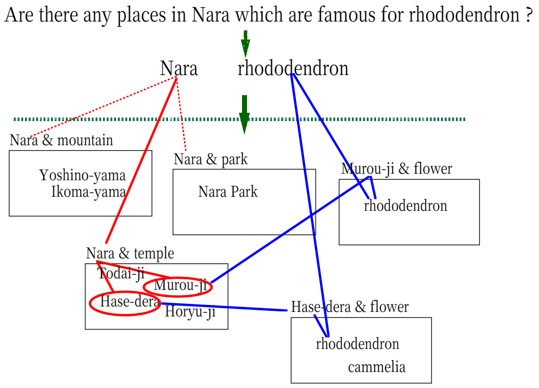


*Fig. 12.* Intelligent Associative Retrieval

### *4.4. Example*

When a user wants to know researchers and their contact information concerning their research interest but there is no such database available, they may search WWW pages and find the information using her/his knowledge. Information integration facility helps the user in such a process.

Fig.7 shows how the facility answers the following question against the sets of 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' ?"

Fig.8 and Fig.9 illustrate example results of the facility.

## 5. Information Retrieval Facility

Information retrieval facility gives users access to 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 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 a selected unit by associations. For example, when an association shown in Fig.2(b) is given and the user selects "AI", linked units such as "people", "Minsky", "McCarthy" and "Weizenbaum" will be displayed. Users can execute neighbor search by pressing buttons displayed nearby units on workspaces [3].

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. Fig.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.

Intelligent associative retrieval is based on the idea of "spreading activation" on semantic networks [3]. The algorithm incrementally extends the neighborhoods of given units and computes the intersection and shown in Fig.11.

### *5.3. Example*

Fig.12 illustrates how the algorithm works to answer a question;

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

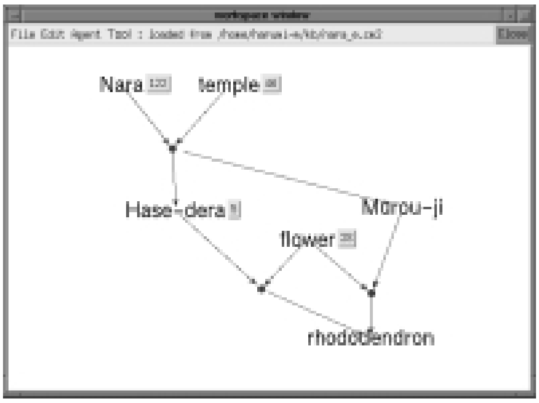


*Fig. 13.* An Example Screen of Intelligent Associative Retrieval

```
; given an information base $B$ and
a threshold $\theta > 0$
repeat
  ⟨use heuristic rules to diagnose $B$ and
  produce a set of suggestions and associated penalties if any undesirable portion
  is found⟩;
  if
    ⟨the largest penalty is greater than
    $\theta$ ⟩;
    then
      ⟨fix $B$ as suggested by the diagnosis
      with the largest penalty ⟩;
    else
      exit from the loop
  for ever
```

*Fig. 14.* A General Procedure of Refinement

Fig.13 shows the result of the intelligent associative retrieval.

In this example, "Hase-dera (temple)" and "Murou-ji (temple)" are extracted because they both are linked to "Nara" and "temple" in the information base.

## 6. Information Refinement Facility

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

Compare two sets of associations in Fig.2(a) and (b). The association of Fig.2(b) is more comprehensible and useful than that of Fig. 2(a), because various kinds of entities are mixed up in the association in Fig.2(a). For understanding the domain and utilizing the information, the latter is more useful.[4]

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

if
  ⟨ for two sets of concepts $\alpha, \beta$, $\alpha \subset \beta$:
       $\exists z[\, z \in V[\alpha] \wedge z \in V[\beta]\, ]$ ⟩
then
  penalty $\leftarrow \infty$ ;
  suggestion $\leftarrow$ "remove $z$ from $V[\alpha]$."
```

*Fig. 15.* Diagnosis Rules for Orthogonal Decomposition

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

### 6.1. *General Procedure*

In order to resolve difficulty with conceptual diversity, we explore two heuristics called Orthogonal Decomposition and Analogical Refinement to suggest the user how to refine the information base into a coherent structure, as shown in Fig14.

### 6.2. *Orthogonal Decomposition*

The orthogonal decomposition searches for a maximal collection of units that have common values in associative structures. For example, there might be a record in the information base saying that "MIT reminds us of Minsky, Winston, Negroponte, Sussman and Maes." Another record may say that "Agents reminds us of Minsky, Negroponte and Maes." The orthogonal decomposition suggests that "MIT and agents remind us of Minsky, Negroponte and Maes."

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:

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

$\mathrm{Sim}^{(a)}[x,y]$ 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:

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

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

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

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

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

*Fig. 16.* Defining Similarity between Concepts

```
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 $\mathrm{V}[\alpha \cup \{a\}]$ .
```

*Fig. 17.* Diagnosis Rules for Analogical Refinement

Figure 15 shows diagnosis rules of the orthogonal decomposition.

### 6.3. *Analogical Refinement*

The analogical refinement extends the orthogonal decomposition by allowing similarity-based matching. The similarity is computed based on the topology of the association network. Roughly speaking, units $x$ and $y$ are regarded as more similar if there are more associations that take both $x$ and $y$ as the value or there are more occasions in which they are orthogonal to the same set of units. For example, the similarity measurement between "Tokyo" and "Kyoto" is increased if one encounters in the information base a record such as "Japanese cities remind us of Tokyo and Kyoto" or "both Tokyo and Kyoto have nonempty intersection with the population, the area, etc."

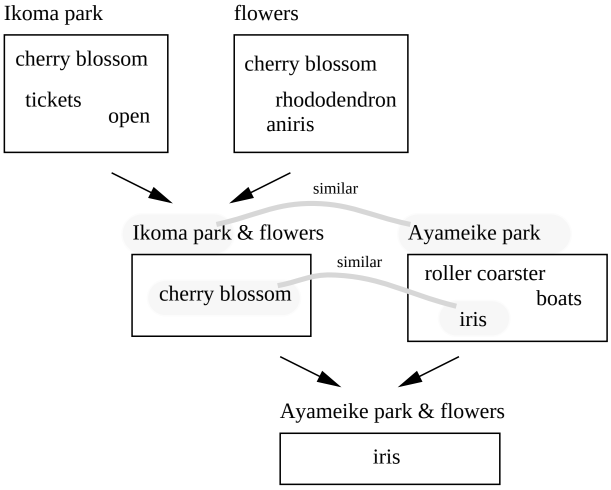


*Fig. 18.* Information Refinement Facility

Given a couple of non-orthogonal keys $x$ and $y$, we define the similarity $\mathrm{Sim}[x,y]$ as shown in Fig.16. Based on that definition, we define the key similarity $\mathrm{Sim}^*[\alpha,\beta]$ between keys $\alpha$ and $\beta$ as the sum of maximal pairwise similarities of units in $\alpha$ and $\beta$. 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 $\alpha$, $\beta$, and a threshold $\theta$, $\alpha \sim \beta$ if $\mathrm{Sim}^*[\alpha,\beta] \geq \theta$.

The analogical refinement heuristic suggests to refine an information base according to the diagnosis rule shown in Fig.17.

### *6.4. Example*

There are several interesting suggestions by the orthogonal decomposition and the analogical refinement as shown in Fig.18.

For example, from "cherry blossom" ∈ V[{"Ikoma park"}] and, "cherry blossom" ∈ V[{"flowers"}], we obtained

"cherry blossom" ∈ V[{"Ikoma park", "flowers"}],

from which we in turn obtained

"iris" ∈ V[{"Ayameike park", "flowers"}]

based on

"iris" ∈ V[{"Ayameike park"}],
"Ikoma park" ∼ "Ayameike park", and
"cherry blossom" ∼ "iris".

## 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.

### *7.1. Experiment 1: Information Capture Facility*

We manually constructed an information base referring a sightseeing guidebook of Nara. It contains 1,315 units and 861 associations. It took 40 hours.

An article of Nikkei Newspaper Database [5] is composed of a header and a content. We implemented capture facility which extracts concepts from headers and external-data from contents. 92% of generated concepts are appropriate as keywords of articles.

### *7.2. Experiment 2: Information Capture & Integration Facility*

We gathered 100 WWW pages concerning AI researchers. CM-2 has extracted units about 7 classes such as researchers, topics and universities and generated associations. 288 heuristics were used to capture the seven classes of "people," "e-mail," "project," "university," "department," "laboratory," and "topic." We have evaluated the result in terms of *precision* and *recall*. Precision means the proportion of correct units over the set of units that information capture facility has generated for the class, while recall means the proportion of actually generated units over the set of potential units created for the class.

We measured the precision and recall for the class "people" (Test 1) and "projects" (Test 2), respectively. Table 1 summarizes the result of the performance evaluation we have made so far. In the case of HTML documents, we obtained 90% as precision and 83% as recall for Test 1, and 68% as precision and 73% as recall for Test 2. The result was worse in the latter, because the original WWW pages are people-oriented and descriptions of projects are relatively subsidiary.

### *7.3. Experiment 3: Information Retrieval Facility*

We tested information retrieval facility against the information base manually constructed from a sightseeing guidebook of Nara in Experiment 1. We asked 50 questions. The performance of the system is evaluated in terms of precision (the ratio of questions resulting in paths containing only appropriate units) and recall (the ratio of questions resulting in paths containing at least one appropriate unit). The result is shown in Table 2. In

*Table 1.* Experimental Results of Intelligent Information Integration Facility

| Test | Precision | Recall |
|---|---|---|
| Test 1 (people) | 90% | 83% |
| Test 2 (project) | 68% | 73% |

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

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

distance 1+2 search, units that cannot be found in a distance 1 search are found. Thus, both precision and recall are increased.

### 7.4. *Experiment 4: Information Refinement Facility*

We tested two heuristics of information refinement facility against the same information base as Experiment 3. Among them, 277 candidates of refinement were proposed, and about 36% of them were judged acceptable by the human evaluator.

## 8. Related Work and Discussion

The work reported in this paper is part of the **Knowledgeable Community** [4] 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 structures.

Our work is related to plenty of search engines on the WWW (e.g. Yahoo [6], AltaVista [7]) and recent work on information gathering from heterogeneous sources on the Internet (e.g. [5],[6],[7],[8]). 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.

*Table 2.* Experimental Results of Information Retrieval Facility

| Test | Precision | Recall |
|---|---|---|
| Test 1 (Distance 1) | 64% | 64% |
| Test 2 (Distance 1 + 2) | 76% | 92% |

Precision: $\frac{answers\ \ containing\ right\ results\ only}{questions} \times 100$ (%)

Recall: $\frac{answers\ \ containing\ right\ results}{questions} \times 100$ (%)

CYC[9] and ARPA Knowledge Sharing Effort[10] have made a significant contribution in the sense they shed light on the importance of knowledge and information sharing and that they have presented a self-completed computational model. Their approach orients computer information sharing, while ours for human information sharing.

Gaines uses semantic networks as information representation for group knowledge sharing[11]. Our approach is based on much weaker information representation than semantic networks.

Concerning information extraction from newspaper articles, many experiments have been done in MUC (Message Understanding Conference) [12]. In MUC, researchers focus on improving the precision of information extraction. In contrast, we present an information representation to integrate heterogeneous information and concentrate on the heuristics of integration.

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 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 structures to accumulate and access varieties of information taken from vast information sources and access relevant information.

However, Experiment 4 shows that there is much scope to improve the heuristics used for information refinement since the rate of useful suggestions from the heuristics seems to be low. To improve the quality of heuristics, we are currently looking at introducing other kinds of heuristics and domain knowledge.

One possibility is introducing a notion of *significance* of association. Given a set of keys $\alpha$, we may ask the user to specify whether $\alpha$ is *significant* or not (*i.e.*, *insignificant*), depending on whether thinking about $\alpha$ makes sense to the user or not, respectively. If the user has explicitly given the contents of $\mathrm{V}^*[\alpha]$, $\alpha$ is taken to be significant. The user may well reserve the remark about the significance of the keys $\alpha$. In such a case, $\alpha$ is called *significance-unknown*. Using this convention, we can avoid thinking about such associations as

“\$1.00” $\in$ V [{“the fare”, “the toll bridge A”, “the price”, “the burger”}].

Or, we might use the following kind of heuristic and store “hindsight” as a more structured conceptual structure.

```
if
    ⟨ $\mathrm{V}^*[\{x,y\}]$ does not make sense, or
    the user assert that
        $\mathrm{V}^*[\{x,y\}] \neq (\mathrm{V}^*[\{x\}] \cap \mathrm{V}^*[\{y\}])$ ⟩
then
    $\forall z[\ z \in (\mathrm{V}^*[\{x\}] \cap \mathrm{V}^*[\{y\}])$
        $\rightarrow \neg\langle z \ isa \ x\rangle \wedge \neg\langle z \ isa \ y\rangle\ ]$.
```

Further research is open for future.

## 9. Conclusions

We proposed an approach based on *knowledge medium* using *associative structures* as a framework of information representation to facilitate raw information from vast information sources to be incorporated without much labor.

We 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 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 discussed the strength and weakness of the method on the analysis of experimental results.

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

## Notes

1. “CM” stands for “Contextual Media” which stands for our long term theoretical research goal.
2. http://www.cs.rochester.edu/u/james/
3. 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.
4. One may complain about fragmentation of information in Fig.2(b) and rather prefer the presentation shown in Fig.2(a). We cope with such complaints by introducing facilities for aggregating information and present it altogether.
5. It contains about 140,000 articles in 1994.
6. http://www.yahoo.com/
7. http://www.altavista.digital.com/

## 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 Twelfth National Conference on Artificial Intelligence (AAAI-94)*,1994.
3. M.R.Quillian, Semantic memory. In Marvin Minsky edition, *Semantic Information Processing*, MIT Press, 1968.
4. 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.
5. 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.
6. 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.
7. 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.

8. 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.
9. R. V. Guha and D. B. Lenat, Enabling Agents to Work Together. In *Communications of ACM*, Vol. 37, No. 7, pages 127–142, 1994.
10. R. S. Patil et al., The DARPA Knowledge Sharing Effort: Progress report. Principles of Knowledge Representation and Reasoning, In *Proceedings of the Third International Conference*, (eds. C. Rich, B. Nebel and W. Swartout), Morgan Kaufmann, 1992.
11. B. R. Gaines and M. L. G. Shaw, Using Knowledge Acquisition and Representation Tools to Support Scientific Communities. In *AAAI-94*, 1994.
12. Grishman, R. and Sundheim, B., Message Understanding Conference -6: A Brief History. In *Proceedings of The 16th International Conference on Computational Linguistics (COLING-96)*, pages 466–471, 1996.

**Harumi Maeda** is a lecturer of the Media Center, Osaka City University. She received her B.A. in psychology from Kyoto University in 1986, M.Sc. in computation from University of Manchester Institute of Science and Technology (UMIST) in 1994, and Doctor of Engineering from Nara Institute of Science and Technology (NAIST) in 1998. She worked for Fujitsu Ltd. as a systems engineer in 1986 - 1995. Her research interests include knowledge representation, human factors in computation and creative thinking support.

**Kazuto Koujitani** is a researcher in the Human Understanding Group at OMRON Corporation. He received his B.E. from Osaka University in 1994, and M.E. from Nara Institute of Science and Technology (NAIST) in 1996. His research interests include knowledge acquisition and spoken dialogue system. He is a member of the Japan Society of Artificial Intelligence and the Association for Natural Language Processing.

**Toyoaki Nishida** is Professor in Artificial Intelligence Laboratory in Graduate School of Information Science at Nara Institute of Science and Technology (NAIST). He received B.E., M.E., and Doctor of Engineering degrees from Kyoto University in 1977, 1979, and 1984 respectively. His research interest covers knowledge media, agent technology, knowledge sharing, and qualitative reasoning. From 1998, he founded the Synsophy project, sponsored by the Japanese government, that aims to study social interaction and community interaction support. He is an area editor (Intelligent Systems) of New Generation Computing, an editor of Autonomous Agents and Multi-Agent Systems, a fellow of FIPA (Foundation of Physical Standardization Agents), and a trustee of JSAI (Japanese Association for Artificial Intelligence).