Ontology-based Approach to Information Gathering and Text Categorization

Michiaki Iwazume, Hideaki Takeda, and Toyoaki Nishida
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama, Ikoma, Nara, 630-01, JAPAN
Phone: +81-07437-9-9211 ext.5316
Fax: +81-07437-2-5269
E-mail: {mitiak-i, takeda, nishida }@is.aist-nara.ac.jp

Abstract

In this paper, we propose a new method of information gathering and text categorization using ontologies. We implemented a system called IICA (Intelligent Information Collector and Analyzer)which helps people to acquire knowledge from information resources on the wide-area network by gathering information and categorizing texts. We tested IICA for (1)gathering pages on the www and(2)categorizing articles on the network news. The results of the experiments indicated that the ontology-based approach enable us to use heterogeneous information resources on the wide-area such as the www and the network news.

Keywords: Ontology, Information gathering, Text categorization, Knowledge media

1 Introduction

Since the number and diversity of information sources on the Internet is increasing rapidly, it becomes increasingly difficult to acquire information we need. A number of tools are available to help people search for the information (for example [6], [5]). However, these tools are unable to interpret the result of their search due to lack of knowledge of the domain. We need more intelligent systems which facilitate personal activities of producing information such as surveying, writing papers and so on.

In this paper, we propose an ontology-based approach to gathering and classifying information in order to realize intelligent agents to help personal activities of information production.

We implemented a system called "IICA" which helps people to acquire knowledge from the information resources on the wide-area network by gathering and categorizing information . Figure 1 shows the outline of IICA.

IICA gathers www pages and USENET network news articles on the Internet in response to user's requests. IICA uses ontologies to compute

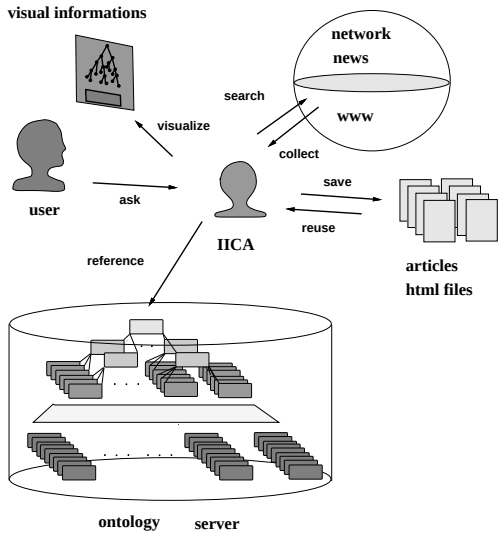

Figure 1: IICA: Intelligent Information Collector and Analyzer

the similarity between the keywords given by the user and those extracted from candidate texts. In case there is no texts which contains the given keywords, IICA infers significant terms related to the given keywords and gathers texts concerned with these terms. Furthermore, IICA categorizes the gathered texts by linking them with an ontology.

We tested IICA for information gathering on the www and text categorization on the network news.

In Section 2, We will first explain the role of ontologies in gathering and categorizing text-like information. We also propose and describe weakly structured ontology which is developed from existing terminologies and thesauruses. In Section 3, we will show how IICA uses ontologies to gather information intelligently, and we will explain a new method of text categorization using ontologies in Section4. In Section 5, we will describe the

```
(con4001
   (representation
      (symbol ``knowledge representation''))
   (context
      ((con4002 (weight 0.75))
       (con4003 (weight 0.80))
       (con4004 (weight 0.30))
                  :                ))))

(con4004
   (representation
      (symbol ``knowledge base'')
   (context
      ((con4017 (weight 0.80))
       (con4018 (weight 1.00))
                  :
       (con4022 (weight 0.50))))))

(con4017
   (representation
       (symbol ``knowledge network'')))
(con4018
   (representation
      (symbol ``knowledge base management'')))
                  :
(con4022
   (representation
      (symbol ``logical database'')))
```

Figure 2: An Example of a Weakly Structured Ontology

implementation of the prototype system of IICA. In Section 6, we will discuss the advantages of our approach and summarize this paper.

2 Ontology

2.1 The role of ontologies

An Ontology is specification of conceptualization which consists of a vocabulary and a theory [2]. The role of ontologies in our approach is fourfold: (a)providing knowledge for agents to infer information which is relevant to user's requests, (b)filtering and classifying information (c)indexing information gathered and classified for browsing , and (d)providing a pre-defined set of terms for exchanging information between human and agents.

2.2 Weakly structured ontologies

Unfortunately, development of ontologies is often a quite painstaking and time consuming task. Ontologies are often described in frame languages such as Ontolingua [1] and knowledge representation languages based on first-order predicate logic. We believe that the difficulty comes from the fact the these languages is computer oriented media and not human-oriented media. Since most of our knowledge is in human media such as natural language documents, we have to somehow translate human-oriented media into computer-oriented media. As human-oriented media is often ill-structured, *i.e.*, ambiguous, indefinite, vague, unstructured, unorganized and inconsistent, we need a tremendous amount of efforts

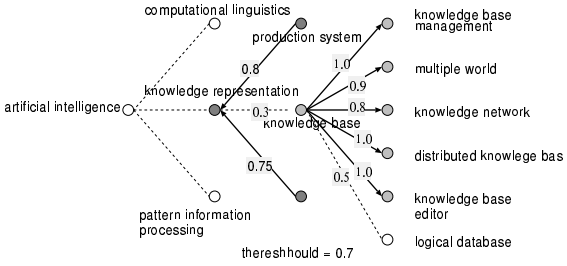

Figure 3: Generating Related Terms to User's Inputs

on translating ill-formed information into well-formed information.

We decided to make use of *weakly structured ontologies* which is developed from existing terminologies, thesauruses [3], and technical books [7]. Weakly structured ontologies have only one type of associative relation between terms. Conceptual relations such as concept-value, class-instance, superclass-subclass, part-whole are not explicitly distinguished in the weakly structured ontologies.

Figure 2 shows an example of a weakly structured ontology. The ontology is described in symbolic expressions. `con4001` is internal ID to an object. `(representation (symbol ''knowledge base''))` means that the external representation of `con4001` is a symbol, "knowledge base". `context` takes an argument of a list of subclass objects. One element in the list consists of an internal ID and a value of `weight` to the subclass object. In the following experiments, we use the ontology built from the information science terminology which has about 4,500 terms.

3 Ontology-based intelligent information gathering

This chapter describes how IICA uses ontologies to gather information intelligently.

3.1 Inference of related terms to user inputs

IICA uses ontologies as shown in Figure 3 to compute the similarity between the keywords given by the user and those extracted from candidate texts.

For example, suppose that a user wants to know information about "knowledge base". In case there is no texts which contains a term "knowledge base", IICA infers that significant terms related to "knowledge base" are not only terms containing the same string like "knowledge base management", "distributed knowledge base", and "knowledge base editor", but also

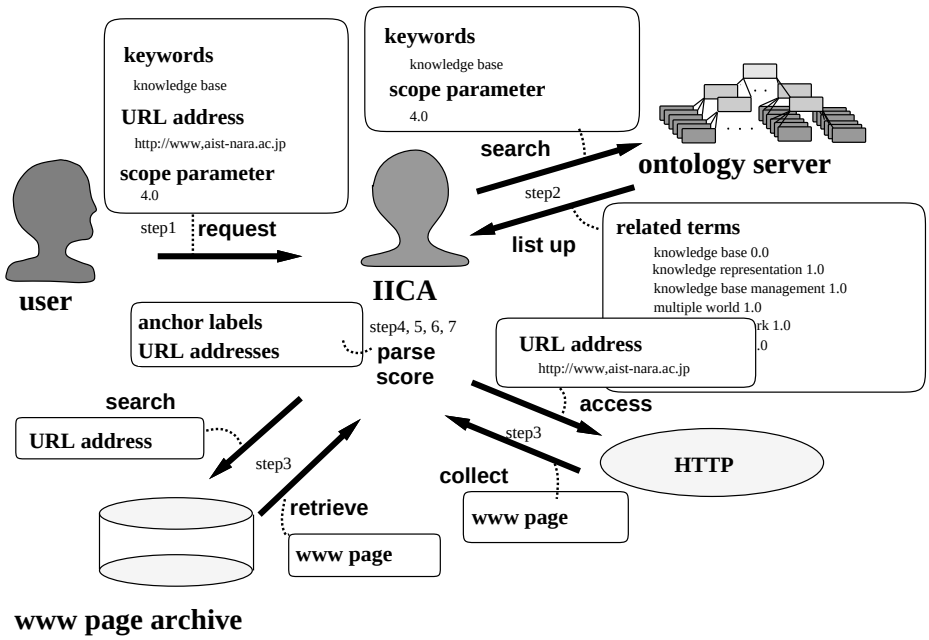

Figure 4: Outline of Information Gathering on the www

terms related ontologically like "knowledge network" and "multiple world". IICA can also reason about the context from user's query. For example, when the input keywords are "semantic network" and "logical database", IICA interprets that the context is "knowledge representation". Inference about the context depends on the level of the terms and the weight values between terms. Users can control the scope of reasoning the context by specifying the threshold parameter.

3.2 Information gathering on the www

In this section, we describe the search algorithm [1] on the www.

Figure 4 is outline of the information gathering on the www. IICA collects pages by (1)accessing HTTP or (2)searching the archive of www pages. In the former case, IICA gets the specified page by sending a URL address to its socket modules and accessing the specified host. The gathered page is added to the archive. All pages in the archive are managed by IICA with its file table . In the latter case, IICA searches the archives using the file table.

3.2.1 Algorithm

The algorithm is basically breadth-first searching. The difference is that IICA evaluates gathered pages and decides which anchor to access next. we show the algorithm as follows.

step1
: Receive a set of keywords, starting URL address, scope of reasoning context and number of pages to gathered from the user.

step2
: Match the keywords with terms in the ontology and list up terms relevant to the within the scope.

step3
: If the specified URL address exists in the close-list, search the page from the archive. Otherwise, retrieve the page by accessing HTTP.

step4
: If the number of pages is greater than the limit, exit the procedure. Otherwise, go to step5.

step5
: Parse the gathered page to extract URL addresses and labels in anchors and titles. If the addresses already exist in the open-list and close-list, discard them. Otherwise, add them to the open-list.

step6
: IF the terms listed up at step2 are included in the labels, score the labels using ontology. Otherwise, remove the label and the

[1] We should take notice that an automatic search on the www often bring about heavy loads on the network. In practical use, some heuristics such as restricting time and frequency to access to the network and avoiding concentrative access to particular hosts is necessary.

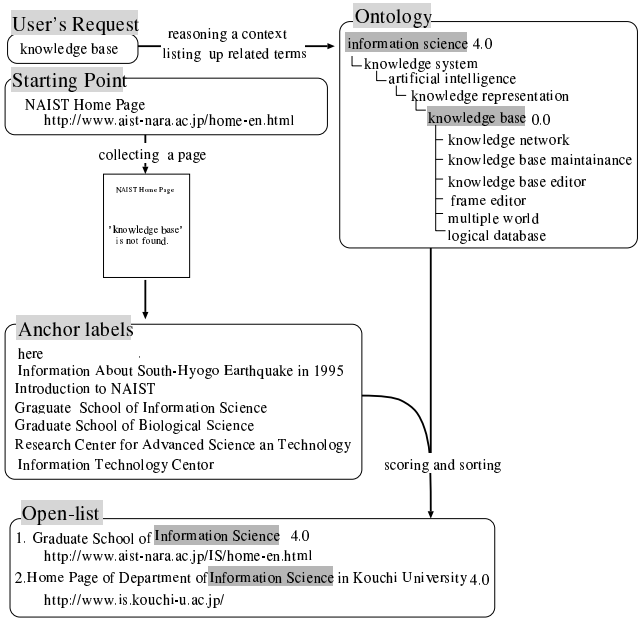

Figure 5: An Example of Information Gathering on the www

addresses from the open-list. Then Sort the open-list.

step7

If there is no anchor in the page, pick up a URL address from the open-list. Then Go to step3.

3.2.2 Example

We describe an example of gathering pages on the www using above algorithm. Suppose that the user's keyword is "knowledge base", the starting URL address is "http://www.aist-nara.ac.jp/home-en.html", and the scope is 4.0 at step1 (see Figure 5). `IICA` generate a set of related terms to the keyword using the ontology (step2). As the specified URL address is not in the open-list or close-list, `IICA` retrieves the page (step3). Moreover, `IICA` extract 27 anchor labels and URL addresses from the page (step5), and score and sort them (step6). In this case, two anchors, "Graduate School of Information Science"(URL address: "http://www.aist-nara.ac.jp/IS/home-en.html", score : 4.0) and "Home Page of Department of Information Science in Kouchi University" (URL address: "http://www.is.aist-nara.ac.jp/", scope: 4.0) are added to the open-list.

3.3 Information gathering on the network news

Gathering articles on the network new is easier than gathering pages on the www, because

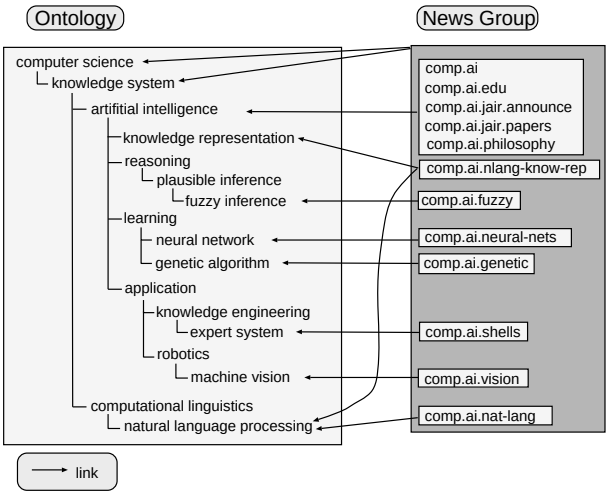

Figure 6: Classification of Newsgroups

the structure of newsgroups is not as complex as that of the www. `IICA` classifys the newsgroups to gather articles efficiency. Figure 6 shows an example of classification of newsgroups.

Since newsgroups are often described in peculiar abbreviations (for example "comp"), it is difficult to match the description of the newsgroups for terms in the ontologies. `IICA` uses heuristics to classify the newsgroup. There are 511 abbreviations in 711 newsgroups of the domain "comp". Figure 7 shows the heuristics which is described in a list of pairs called an *association list*. The *car* of a pair is an abbreviation, and the *cdr* is an ontological term. It is not necessary to give heuristics to every newsgroup, because abbreviations are used in newsgroups of other domain such as "alt".

```
("comp" . "computer science")
("ai" ."artificial intelligence")
("edu" . "education")
("fuzzy" . "fuzzy inference")
("genetic" . "genetic algorithm")
("nat-lang" . "natural language processing")
("neural-nets" . "neural network")
("nlang-know-rep" . "natural language processing")
("nlang-know-rep" . "knowledge representation")
("shells" . "expert system")
("vision" . "machine vision")
("infosystems" . "information system")
```

Figure 7: Heuristics for Classification of Newsgroups

3.4 Experiment

We chose a query which consists of a keyword "knowledge base" , a starting URL ad-

Table 1: Result of Information Gathering on the www

generated ontological terms	term frequency	scope
knowledge	461 (92.2 %)	2.0
knowledge base	451 (90.2 %)	0.0
artificial intelligence	425 (85.0 %)	2.0
production system	252 (50.4 %)	2.0
knowledge representation	101 (20.2 %)	1.0
distributed knowledge base	90 (18.0 %)	1.0
blackboard model	72 (14.4 %)	2.0
semantic netwrok	40 (8.0 %)	2.0
multiple world	26 (5.2 %)	1.0
knowledge network	11 (2.2 %)	1.0
predicate logic	10 (2.0 %)	2.0
frame	10 (2.0 %)	2.0
knowledge base management	2 (0.4 %)	1.0
knowledge base editor	0 (0.0 %)	1.0
logical database	0 (0.0 %)	1.0
inference control	0 (0.0 %)	2.0
representation of ambiguity	0 (0.0 %)	2.0

dress "http://ai-www.aist-nara.ac.jp" scope parameters "2.0", and ran `IICA` on the www. `IICA` gathered 500 pages for about 12 hours using DEC 3000 alpha AXP. Table 1 shows generated a set of terms related to the input "knowledge base", term frequency in the collection of the pages and scope parameters. 90.2 % of collected pages correctly contain keyword "knowledge base".

4 Ontology-based text categorization

Ontology-based text categorization is the classification of documents by using ontologies as category definition. Conventional approaches focused only on the accuracy of categorization and left the easiness of human understanding out of consideration. Our purpose is extending the conventional methods using ontologies. Ontologies help people to interpret the result of categorizing texts by showing the ontological relations between texts.

In our approach, the process of text categorization is twofold: (1)Calculating similarity between a feature vector and a category vector, (2)Calculating similarity between category vectors (see Figure 8).

A feature vector is a vector which represents feature of a document. The feature vector is calculated from the term frequency and the inverse document frequency *A category vector* is a vector which represents the characteristic of a category. The category vector is calculated from the feature vectors of the document assigned to the category.

4.1 Text categorization using structured knowledge

There are studies on text categorization using structured knowledge such as thesaurus[4] [9]. However, in these approaches, it is difficult to deal with changeable information resources, because a link between terms in the thesaurus is fixed, and category vectors are strongly depended on the initial learning data. Moreover, it is impossible to retrieve texts in categories similar to the current category, because there is no consideration of similarity between categories.

In our approach, it is possible to use actual information resources by modifying not only category vectors dynamically but also weight between categories from gathered data. Furthermore, it is impossible to retrieve texts in categories similar to the current category by calculating similarity between categories. Category vectors and weights are calculated as follow procedures.

step1
: Calculate the feature vectors of the gathered text.

step2
: Classify gathered texts by calculated the feature vector.

step3
: Calculate the category vectors from the classified texts.

step4
: Repeat step2 and step3 until the category vectors converge.

step5

Calculate distance between the categories and renew weight between terms in the ontology.

The each initial category vectors is calculated from the feature vector of the texts which is assigned to the category by matching keywords.

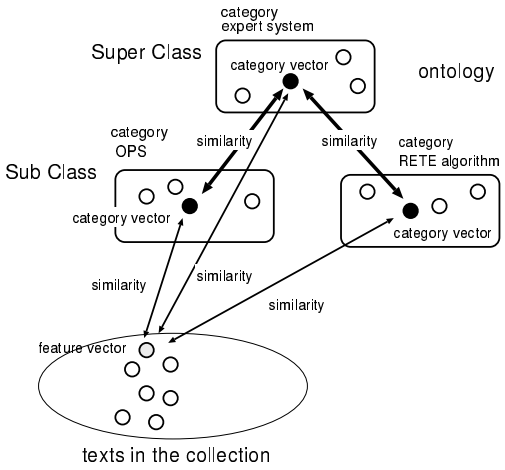

Figure 8: Text Categorization Using an Ontology

4.2 Vector space model

We use vector space model commonly used in the information retrieval studies to weight terms and calculate feature vectors [8].

The weight of term is a product of its term frequency (tf) and its inverse document frequency (idf).

The tf is the occurrence frequency of the term in the text. It is normally reflective of term importance.

The idf is a factor which enhances the terms which appears in fewer documents, while downgrading the terms occurring in many documents. It means that the document-specific feature are highlighted, while the collection-wide feature are diminished in importance. The weight of the term is given as

$$w_{ik} = tf_{ik} \times idf_k,$$

where tf_{ik} is the number of occurrences of term t_k in document i, and idf_k is the inverse document frequency of the term t_k in the collection of documents. A commonly used measure for the inverse document frequency is

$$idf_k = log(N/n_k),$$

where N is the total number of documents in the collection, and n_k is the number of document which contains a given term t_k. The collection of documents is the context within which the inverse documents frequencies are evaluated.

Table 2: Result of Classifying Articles

the top 20 categories and the number of texts		the low 20 categories and the number of texts	
program	48	VLSI	1
planning	31	statistics	1
aritificial intelligence	25	SQL	1
prolog	17	signal	1
software	16	psycology	1
inference engine	14	PC	1
classification	13	lisp	1
cognitive science	12	interface	1
expert system	10	informatics	1
C	9	DOS	1
Turing	8	device	1
neural network	8	design	1
TSP	7	connectionism	1
information	7	computer security	1
concept	7	compiler	1
communication	6	chess machine	1
search	6	brain	1
fuzzy	6	bag	1
IEEE	6	backpropagation	1
backtracking	6	analog computer	1

4.3 Experiment

We tested the above procedure by categorizing 400 articles about "artificial intelligence" on the USENET network news. We chose newsgroups "comp". `IICA` classified the articles to 75 categories. Table 2 shows a part of the results. The table in the left-hand side shows the highest 20 categories and the number of articles and the table in the right-hand side is the lowest 20 categories and the number of articles.

Table 3: Evaluation of Classifying Articles

Accuracy(%)	Recall(%)	Precision(%)
77.0	76.2	76.0

In order to evaluate the result, we calculated Accuracy(A), Recall(R) and Precision(P) using the following equations:

$$Accuracy = \frac{No.\ of\ texts\ assigned\ to\ the\ correct\ category}{No.\ of\ total\ texts\ \ in\ the\ collection},$$

$$Recall = \frac{No.\ of\ texts\ assigned\ to\ the\ correct\ category}{No.\ of\ total\ texts\ \ in\ the\ category},$$

$$Precision = \frac{No.\ of\ texts\ assignedto\ the\ correct\ category}{No.\ of\ total\ texts\ assigned\ to\ the\ category}.$$

The result of calculation is shown in Table 3 , where values of "Recall" and "Precision" are the average of all categories. We also analyzed misclassifications and discriminated them to 3 groups. The first group contains cases in which texts are assigned to the subclass of the correct category. The second group contains cases in which texts are assigned to the superclass of the correct category. And, the third group contains

Table 4: Groups of Misclassifications

result type	No. of texts
texts assigned to the subclass category	26
texts assigned to the superclass category	5
texts assigned to the other category	51

cases in which texts are assigned to the unrelated classes with the correct category. The result of the analysis is shown in Table 4.

Table 5: Revaluation of the Experiment

Accuracy(%)	Recall(%)	Precision(%)
85.3	85.1	85.1

Misclassification of the first and second groups is not serious, because the user can access the misclassified items by tracing ontological relation between categories. Table 5 shows revaluation of the experiment regarding the two groups as correct. In conventional approaches, the misclassified items are not accessible. In contrast, IICA allows the user to search and reach the items by using ontological relations.

5 IICA: Intelligent Information Collector and Analyzer

We implemented a prototype system of "IICA". IICA consists of fourfold modules: (1)user interface modules, (2)network modules , (3)inference modules, (4)database modules. User interface modules is described in Tck/Tk and perl scripts. Network modules is socket programs in C. Inference modules and database modules is implemented in Common Lisp. Figure 9 shows the interface the system using NCSA Mosaic.

IICA also has an ontological browsing tool realized as a *knowledge medium* which unifys the human oriented media and the computer oriented media (See Figure 10). It helps nonprofessional users to search for information and understand the result of categorizing them by visualizing the ontological structures.

6 Conclusion

In this paper, we proposed a new method of information gathering and text categorization using ontologies.

We implemented a system called "IICA (Intelligent Information Collector and Analyzer)"

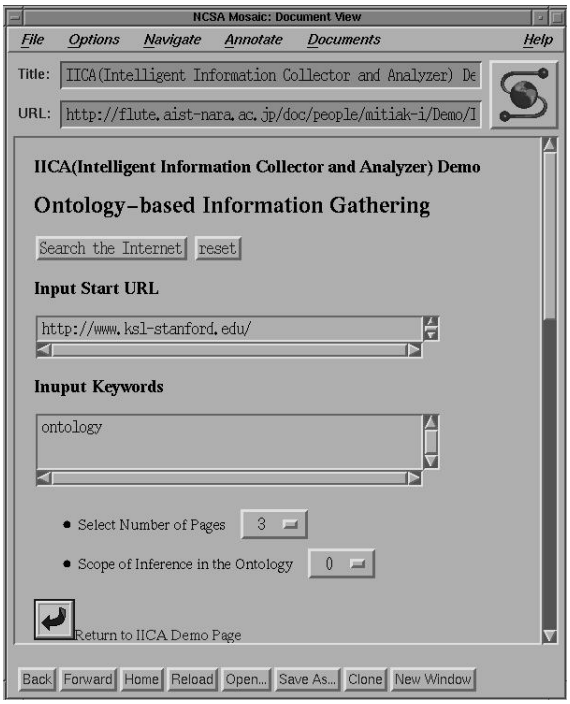

Figure 9: Interface of IICA

which helps people to acquire knowledge from the information resources on the wide-area network gathering and categorizing information.

IICA can deal with various types of text-like information, because most of our knowledge we can use are described as text form.

We tested our approach for two experiments: (1)gathering pages on the www, (2)categorization articles on the network news.

We can conclude the following advantages of our approach from the results of the experiments.

- The ontology-based approach enable us to use heterogeneous information resources on the wide-area such as the www and the network news.
- Agent can understand which information is related to user's request using ontologies.
- In conventional approaches, the misclassified items are not accessible. In contrast, IICA allows the user to search and reach the items by using ontological relations.
- It is easer to develop weakly structured ontologies from terminologies and thesauruses than conventional methods.

The problem of the current system is that ontologies it uses are given and therefore not flexible both to users and information. We should consider learning of new terms from gathered texts and customizing of ontologies to user's interest and purposes.

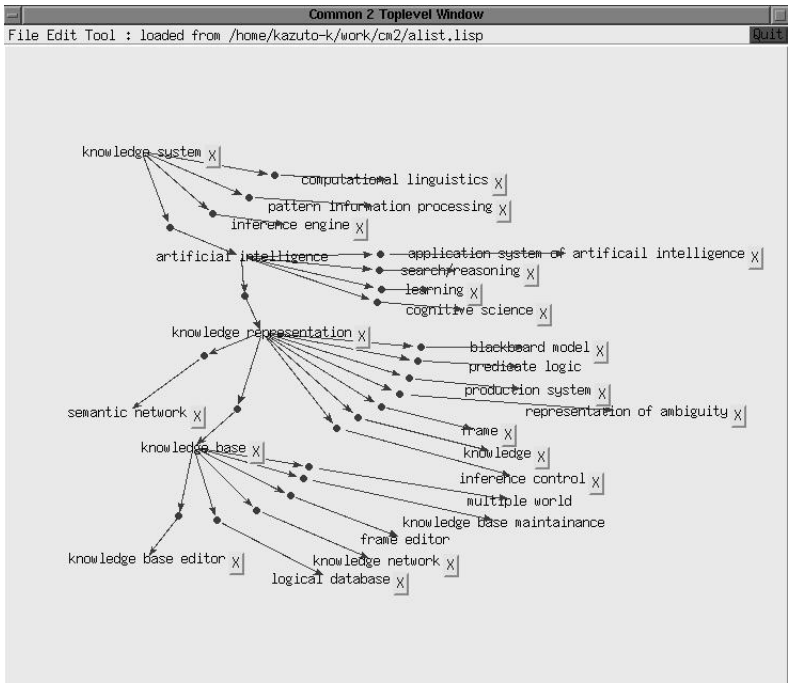

Figure 10: Visualization of the Ontological Structures

References

[1] T. R. Gruber. Ontolingua: A mechanism to support portable ontologies. Technical Report KSL 91-66, Stanford University, Knowledge Systems Laboratory, 1992.

[2] Thomas R. Gruber. The role of common ontology in achieving sharable, reusable knowledge bases. In J. A. Allen, R. Fikes, and E. Sandewell, editors, *Principles of Knowledge Representation and Reasoning – Proceedings of the Second International Conference*, pages 601–602. Morgan Kaufmann, 1991.

[3] Michiaki Iwazume, Hideaki Takeda, and Toyoaki Nishida. Automatic classification of articles in network news and visualization of discussions — intelligent news reader. *Proceedings of the 8th Annual Conference of JSAI, 1994*, 1994.

[4] Atsuo Kawai. An automatic document classification method based on a semantic category frequency analysis. *Transactions of Information Processing Society of Japan*, 33(9):1114–1122, 1992.

[5] P. Maes and R. Kozierok. Learning interface agents. *AAAI-93*, pages 459–465, 1994.

[6] O. McBryan. Genvl and wwww:tools for taming the web. In *Proc. 1st Int. WWW Conf.*, 1994.

[7] Masanobu Nishiki, Hideaki Takeda, and Toyoaki Nishida. Extraction, unification and presentation ok knowledge by multi agent system. *Proceedings of the 8th Annual Conference of JSAI, 1994*, 1994.

[8] G Salton. Intoroduction to modern infomation retrieval. *MacGraw-Hill*, 1983.

[9] Kazuhide Yamamoto, Shigeru Masuyama, and Shuzo Maito. An automatic classification method for japanese texts using mutual category relations. *IPSJ SIG Notes*, 95(27):7–12, 1995.