ONTOLOGY-BASED INFORMATION GATHERING AND CATEGORIZATION FROM THE INTERNET

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
Email: {mitiak-i, takeda, nishida}@is.aist-nara.ac.jp

Abstract

In this paper, we propose a new method to develop more intelligent navigation system for the Internet using ontologies. We implemented a system called IICA (Intelligent Information Collector and Analyzer)which helps people to acquire knowledge from information resources on the Internet by gathering and categorizing information. We tested IICA for tasks on the World Wide Web (WWW) and 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, the WWW, Network news, 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 [McBryan 94], [Maes 93]). However, these tools are unable to interpret the result of their search due to lack of knowledge. 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 Internet by gathering and categorizing information . Figure 1 shows the outline of IICA.

The function of IICA is twofold. (1) Information Gathering: IICA gathers WWW pages and USENET network news articles on the Internet in response to user's requests. IICA uses ontologies to compute the

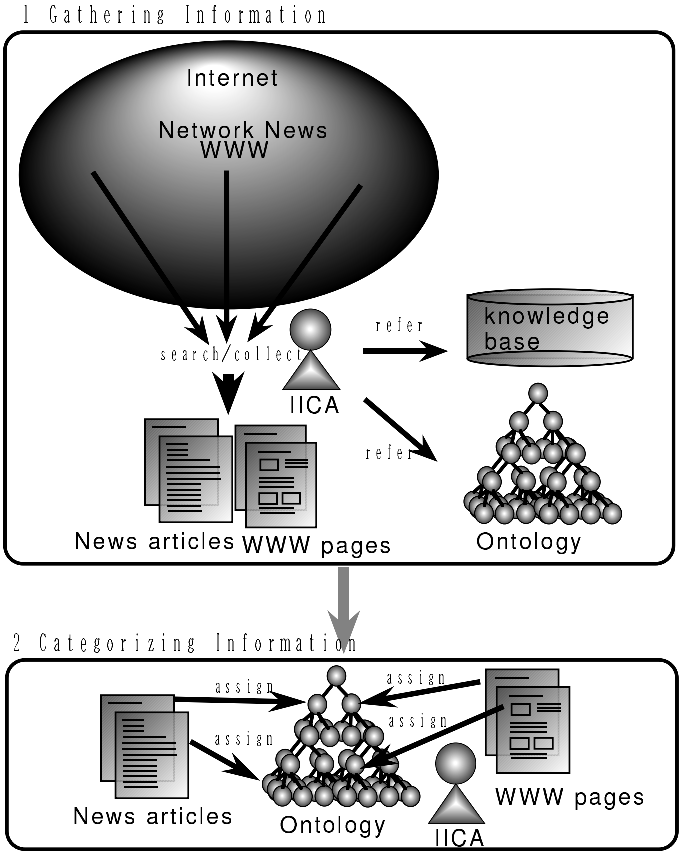


Figure 1: IICA: Intelligent Information Collector and Analyzer

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. (2) Information Categorizing: Furthermore, IICA categorizes the gathered texts by linking them with an ontology. The system helps nonprofessional users to search for information and understand the result of categorizing them by visualizing the ontological structures.

We tested IICA for tasks on the World Wide Web (WWW) and the network news. The results of the experiments indicated that the ontology-based approach enable us to use heterogeneous information resources on the Internet.

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 and heuristics 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 evaluation of the above two methods. In Section 6, we will explain the implementation of the prototype system briefly. In section 7, we will compare our work with related work. In section 8, we will discuss the advantage of our approach and summarize this paper.

2 Ontology

2.1 The Role of Ontology

An Ontology is specification of conceptualization which consists of a vocabulary and a theory [Gruber 91]. 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 Ontology

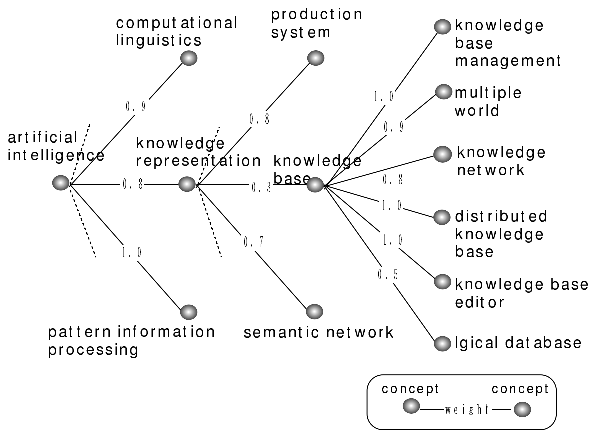


Figure 2: An Example of a Weakly Structured Ontology

Unfortunately, development of ontologies is often a quite painstaking and time consuming task. Ontologies are often described in frame languages such as Ontolingua [Gruber 92] 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 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 [Iwazume 94], and technical books [Nishiki 94]. 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.

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.

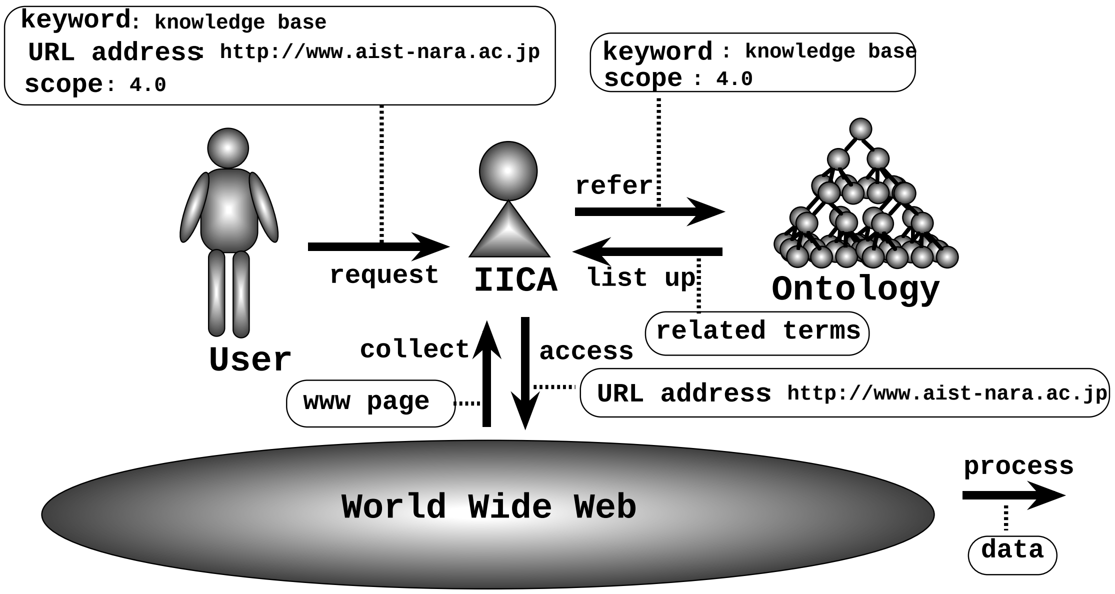


Figure 3: Outline of Information Gathering on the WWW

3.1 Inference of Related Terms to User Inputs

IICA uses ontologies 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 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 3 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.

[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.

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 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 4). `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.

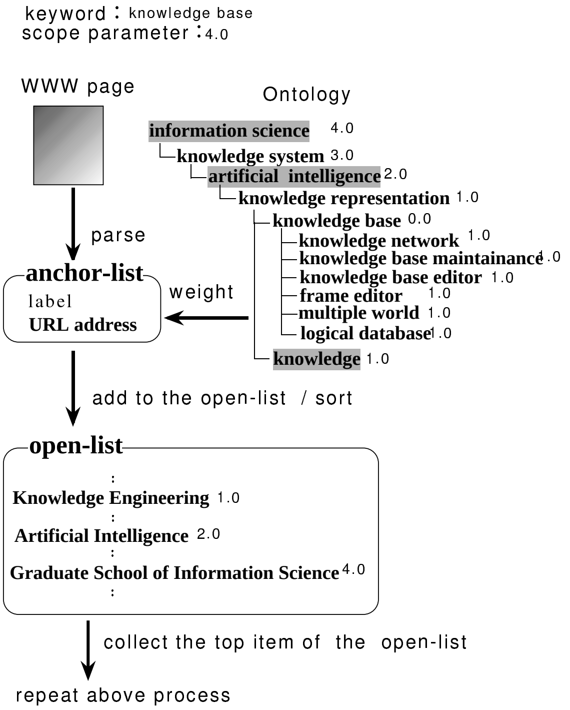


Figure 4: An Example of Information Gathering on the WWW

3.2.3 Heuristics

We often use various heuristics such as empirical knowledge and common sense, when we search for information on the WWW. For example, it is better for us to search for information about AI(artificial intelligence) using a heuristic that

"*the WWW page of institutes, laboratory often contains information about AI*".

We decided to make use of heuristics. For instance, the heuristic that "*if search for information on AI, go pages of laboratory*" is described as follows:

```
‘‘artificial intelligence’’ → ‘‘laboratory’’
```

`IICA` gives priority over the pages which contain term *"laboratory"* and access them by using the heuristic.

```
‘‘artificial intelligence’’ → ‘‘laboratory’’
‘‘artificial intelligence’’ → ‘‘institute’’
```

3.3 Information Gathering on the Network News

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

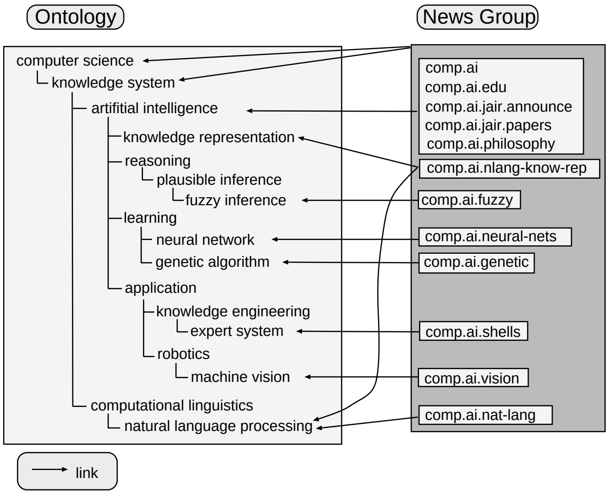


Figure 5: Classification of Newsgroups

of newsgroups is not as complex as that of the WWW. IICA classifies the newsgroups to gather articles efficiency. Figure 5 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 6 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 6: Heuristics for Classification of Newsgroups

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 7).

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 Algorithm

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.

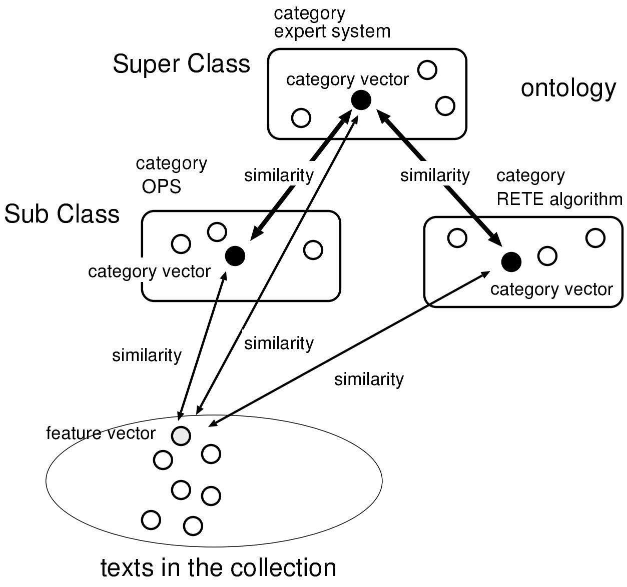


Figure 7: Text Categorization Using an Ontology

4.2 Vector Space Model

We use vector space model commonly used in the information retrieval studies to weight terms and calculate feature vectors [Salton 83].

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.

5 Evaluation

This chapter describes evaluation of our method.

5.1 Evaluation of Information Gathering

We tested an ontology-based method for information gathering tasks on the WWW. We evaluated our system by accuracy and efficiency.

5.1.1 Test of Accuracy

In order to evaluate its accuracy, we restricted 100 pages, and chose the 5 queries related to AI in English and the 5 queries related to sightseeing in Japanese. Then we ran IICA on the WWW in the following ways.

1.Breadth first search: IICA does't use ontologies. It traces hyperlinks on the WWW using breadth first algorithm.

2.Ontology based search: IICA use ontology-based search algorithm.

3.Ontology based search and heuristics: IICA use ontology-based search algorithm and heuristics.

We evaluated the result of the experiment according to the standard as follows.

○: The collected page is directly related to user's queries.

△: The collected page is not directly related to user's queries, but it is related to user's interests.

×: The collected page is neither directly related to the user's queries nor related to user's interests.

Table 1 and Table 2 shows the results.

Table 1: Evaluation of Gathering Pages Relevant to Artificial Intelligence

search	○ (%)	△ (%)	× (%)
1. breadth first search	64.6	7.4	28.0
2. ontology	66.6	11.6	21.8
3. ontology and heuristics	67.8	10.6	21.6

Table 2: Evaluation of Gathering Pages Relevant to Sightseeing

search method	○ (%)	△ (%)	× (%)
1. breadth first search	57.4	8.4	34.2
2. ontology	59.5	15.6	24.9
3. ontology and heuristics	59.5	15.6	24.9

Table 3: Evaluation of Efficiency of Information Gathering — 1 keyword ("knowledge base")

search method	○	△	×
1. bread first search	3	3	3
2. ontology	21	8	12
3. ontology and heuristics	44	13	25

5.1.2 Test of Efficiency

We tested search efficiency of our method. We restricted 500 search steps and chose the 2 queries related to AI in English. Then we ran IICA on the WWW in the above three ways.

Table 3 shows the search result to the query consists of one keyword *"knowledge base"*. Table 4 shows the search result to the query which consists of two keywords *"semantic network"* and *"production system"*. Here, the numbers in this Table indicate number of pages.

5.2 Evaluation of Text Categorization

5.2.1 Experiment on the Network News

We tested our method by categorizing 400 articles about "artificial intelligence" on the USENET network news. We chose newsgroups "comp". IICA classified the articles to 75 categories. Table 5 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.

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

$$A = \frac{No.\ of\ texts\ assigned\ to\ the\ correct\ category}{No.\ of\ total\ texts\ \ in\ the\ collection},$$
$$R = \frac{No.\ of\ texts\ assigned\ to\ the\ correct\ category}{No.\ of\ total\ texts\ \ in\ the\ category},$$
$$P = \frac{No.\ of\ texts\ assignedto\ the\ correct\ category}{No.\ of\ total\ texts\ assigned\ to\ the\ category}.$$

Table 4: Evaluation of Information Gathering — 2 keywords ("semantic network" and "production system")

search method	○	△	×
1. breadth first search	0	0	0
2. ontology	10	12	11
3. ontology and heuristics	18	23	15

Table 5: 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 13	PC	1
classification	12	lisp	1
cognitive science	10	interface	1
expert system	9	informatics	1
C	8	DOS	1
Turing	8	device	1
neural network	7	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

The result of calculation is shown in Table 6 , 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 cases in which texts are assigned to the unrelated classes with the correct category. The result of the analysis is shown in Table 7.

Misclassification of the first and second groups is not serious, because the user can access the misclassi-

Table 6: Evaluation of Classifying Articles

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

Table 7: 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

Table 8: Revaluation of the Experiment

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

fied items by tracing ontological relation between categories. Table 8 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.

5.2.2 Experiment on the WWW

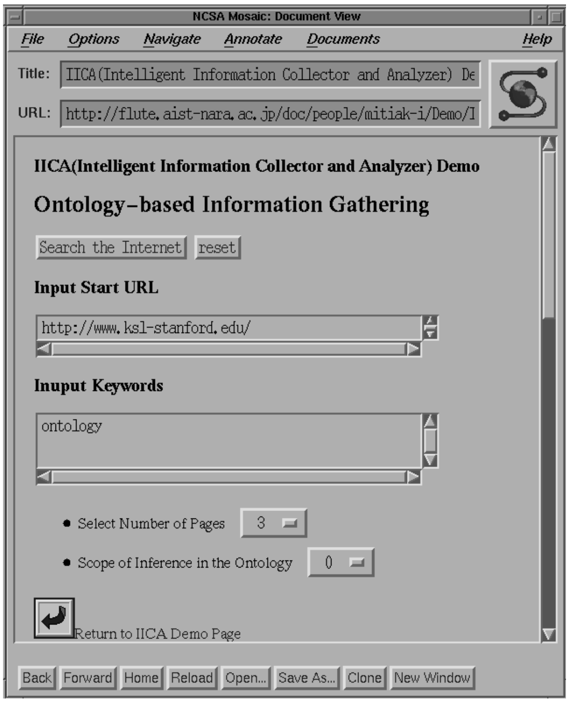
We also made an experiment of categorizing the about 500 pages concerned with AI in English and the about 800 page concerned with sightseeing in Japanese. In order to evaluate our method, we calculated recall and precision. The result is shown in Table 9.

Table 9: Evaluation of Categorization of WWW pages

	AI (English)	Sightseeing (Japanese)
Precision	81.9	79.0
Recall	80.5	70.0

6 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 8 shows the interface the system using NCSA Mosaic.



Title: IICA(Intelligent Information Collector and Analyzer) De

URL: http://flute.aist-nara.ac.jp/doc/people/mitiak-i/Demo/I

IICA(Intelligent Information Collector and Analyzer) Demo

Ontology-based Information Gathering

Input Start URL

http://www.ksl-stanford.edu/

Inuput Keywords

ontology

- Select Number of Pages 3
- Scope of Inference in the Ontology 0

Return to IICA Demo Page

Figure 8: Interface of IICA

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 9). It helps nonprofessional users to search for information and understand the result of categorizing them by visualizing the ontological structures.

7 Related Work

7.1 Internet Robots and Agents

There has been much recent work whixh is concerned with building an index of information on the Internet [Mauldin 94], [Balabanovic 95]. Another related area work attempts to automatically filter incoming information [Lashkari 94]. However, these tools are unable to interpret the result of their search due to lack of dmain knowledge.

In our approach, ontology-based interface helps nonprofessional users to search for information on the Internet and understand the result of categorizing them by visualizing. Ontology also make accurcy and efficiency better in information gathering.

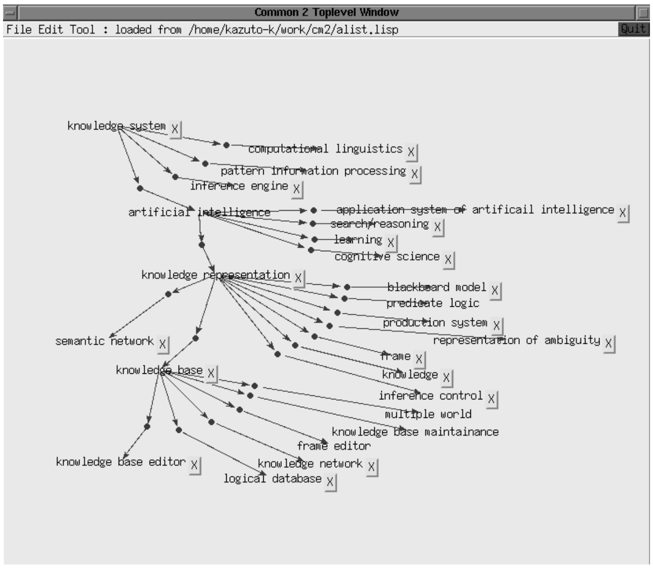


Figure 9: Visualization of the Ontological Structures

7.2 Information Retrieval and Text Categorization

There are studies on text categorization using structured knowledge such as thesaurus [Kawai92e] [Yamamoto95]. 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 the field of information retrieval, the approach based on using Kohonen's self-organizing map have been received attention [Kohonen 90], [Niki 95].

The merit of this approach is providing a collection ambiguous-query interface or an incremental browser for text databases by mapping a collection of documents into a two-dimensional map without teaching. However, it is often difficult for users to understand the meaning of the structure of the map which is organized from raw data.

In out 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.

8 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)" 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 news articles and WWW pages .

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

- Ontology and heuristics make accuracy and efficiency better in information gathering.
- Agent can understand which information is related to user's request using ontologies.
- IICA allows the user to search and reach the the misclassified items by tracing ontological relations.
- It is easier to develop weakly structured ontologies from terminologies and thesauruses than conventional methods.
- The ontology-based approach enables us to use heterogeneous information resources on the Internet.

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.

References

[McBryan 94] O. McBryan, "GENVL and WWWW:Tools for taming the Web", *In Proceedings of the 1th International WWW Conference*, 1994.

[Maes 93] P. Maes and R. Kozierok, "Learning Interface Agents", *In Proceedings of AAAI*, 1993.

[Gruber 91] T. R. Gruber, "The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases", *Principles of Knowledge Representation and Reasoning – Proceedings of the 2nd International Conference,* pp. 601–602, 1991.

[Gruber 92] T. R. Gruber, "Ontolingua: A mechanism to support portable ontologies", *Technical Report of Stanford University, Knowledge Systems Laboratory,* col. KSL 91-66, 1992.

[Iwazume 94] , Michiaki Iwazume and 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,* pp. 497–500, 1994.

[Iwazume 95] Michiaki Iwazume, "Classification and organization of infomation by ontology", *Master's Thesis, Department of Infomation Processing, Graguate School of Infomation Science, Nara Institute of Science and Technology,* vol. NAIST-IS-MT351014, 1995.

[Nishiki 94] Masanobu Nishiki and Hideaki Takeda and Toyoaki Nishida, "Extraction, Unification and Presentation ok Knowledge by Multi Agent System", *Proceedings of the 8th Annual Conference of JSAI,* pp. 505–508, 1994.

[Salton 83] G. Salton, *Intoroduction to Modern Infomation Retrieval*, MacGraw-Hill, 1983.

[Mauldin 94] M. L. Mauldinand and J. R. Leavitt, "Web-agent related research at the CMT", *In Proceedings of the ACM Special Interest Group on Networked Information Discovery and Retrieval*, 1994.

[Balabanovic 95] , M. Balabanovic, "Leaning Information Retrieval Agents: Experiments with Automated Web Browsing", *Proceedings of the AAAI Spring Symposium,* pp. 13–18, 1995.

[Lashkari 94] Y. Lashkari and M. Metral and P. Maes, "Collaborative interface agents", *In Proc of the 12th National Conference on Artificial Inteligence,* 1994.

[Yamamoto95] Kazuhide Yamamoto and Shigeru Masuyama and Shuzo Maito, "An Automatic Classification Method for Japanese Texts using Mutual Category Relations", *IPSJ SIG Notes*, vol. 95, No. 27, pp. 7–12, 1995.

[Kawai92e] Atsuo Kawai, "An Automatic Document Classification Method Based on a Semantic Category Frequency Analysis", *Transactions of Information Processing Society of Japan,* vol. 33, No. 9, pp. 1114–1122 1992.

[Kohonen 90] T. Kohonen, "The Self-Organizing Map", *Proceedings of the IEEE,* vol. 78, No. 9, pp. 1464–1480, 1990.

[Niki 95] Niki Kazuhisa and Katsumi Tanaka, "Information Retrieval Using Neural Networks", *Journal of Japanese Society for Artificial Intelligence,* vol. 10, No. 1, pp. 45–51, 1995.