IICA: An Ontology-based Internet Navigation System

Michiaki Iwazume, Kengo Shirakami, Kazuaki Hatadani
Hideaki Takeda, and Toyoaki Nishida
phone: +81-7437-2-5265, fax: +81-7437-2-5269
{mitiak-i, kengo-s, kazuak-h, takeda, kenji-i, nishida}@is.aist-nara.ac.jp
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama, Ikoma, Nara 630-01, Japan

Abstract

In this paper, we present a system called IICA (Intelligent Information Collector and Analyzer) which gathers, classifies, and reorganizes information from the Internet. Ontology plays an important role in IICA. It specifies the common background knowledge shared by the user and IICA, allows IICA to make inexact match between the user's request and the candidates, and assigns user-oriented categories. IICA extracts information using a state transition network grammar and concept frames. We have implemented and evaluated IICA. The results shows the feasibility and robustness of the approach.

Keywords: Ontology, Internet, WWW, information gathering, information classification, information extraction and reorganization

Introduction

As the number and diversity of information sources on the Internet is increasing rapidly, there is an increase demand for intelligent assistants which would help people search for desired information.

A number of tools are available to help people search for information on the Internet such as *WWW Worm* (McBryan 1994), *WebCrawler* (Pinkerton 1994) Unfortunately, existing tools are unable to interpret the content of information resources due to the 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 present IICA which gathers, classifies, and reorganizes information from heterogeneous information resources on the Internet. Ontology plays an important role in IICA. It specifies the common background knowledge shared by the user and IICA, allows IICA to make inexact match between the user's request and the candidates, and assigns user-oriented categories. Figure 1 shows the outline of IICA.

This system has the following functions. (1) Information Gathering: IICA gathers WWW pages on the Internet in response to user's requests. IICA uses ontologies to compute the similarity between the keywords given by the user and those extracted from candidate pages. (2) Information Categorizing: IICA categorizes the gathered pages by linking them with an ontology and (3) Information Reorganizing: IICA extracts specific information form pages using heuristics based on expression patterns and phrases (See Figure 2).

We tested IICA on the WWW. The results of the experiments suggests that the ontology-based approach enables us intensive use of heterogeneous information resources on the wide-area networks such as the Internet. In Section 2, We describe ontology for information gathering, categorization and reorganization. In Section 3, We explain an information gathering method using ontologies and heuristics. In Section 4, we explain a new method of text categorization using ontologies. In Section 5, we describe how IICA uses heuristics based on expression patterns and phrases to extract and reorganization specific information from pages. In Section 6, we describe the evaluation of the above three methods. In Section 7, we discuss the advantages of our approach and summarize this paper.

Ontology

The Role of Ontology

An Ontology is specification of conceptualization which consists of a vocabulary and a theory (Gruber 1991). 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.

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 1992) and knowledge representation languages based on first-order predicate logic. We believe that the difficulty comes from the fact the these

From: AAAI Technical Report WS-96-06. Compilation copyright © 1996, AAAI (www.aaai.org). All rights reserved.

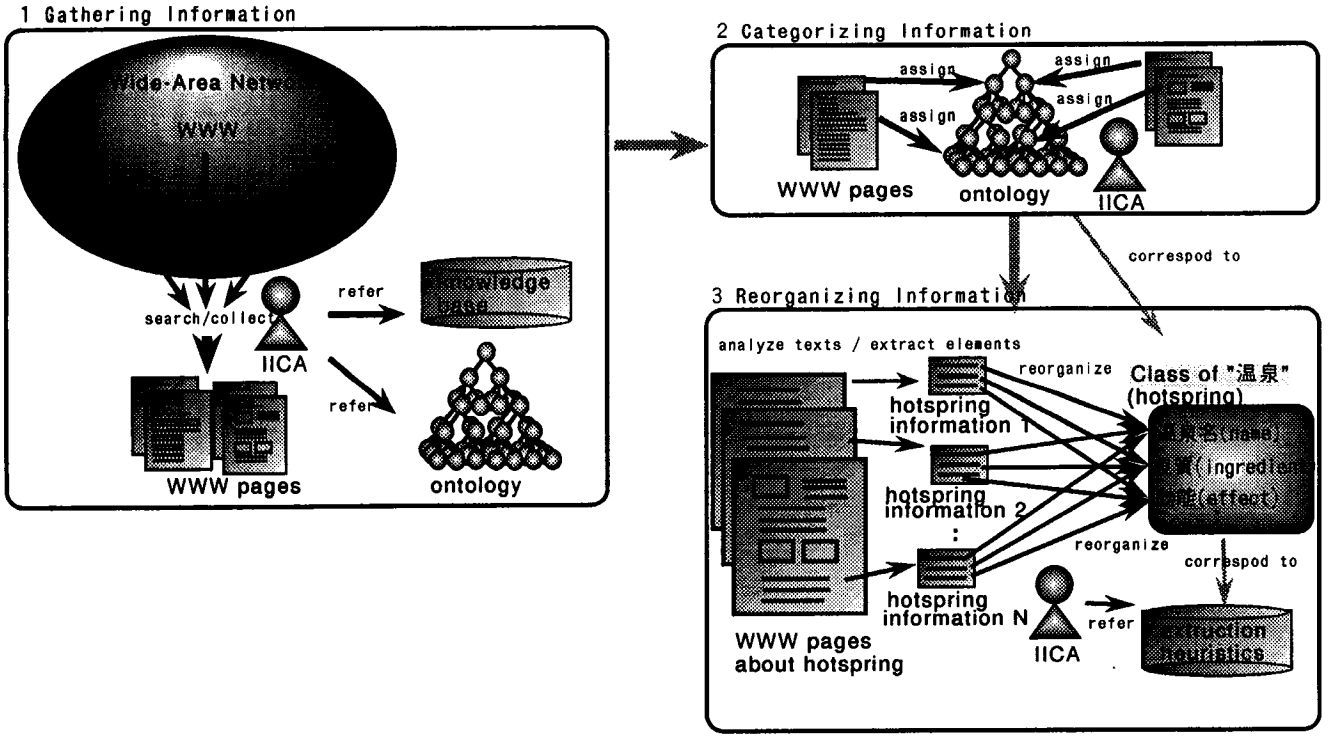

Figure 1: Outline of IICA

URL	温泉の名前	最寄り駅	アクセス方法	風呂の種類	泉質
akase-spa-j.html	"赤瀬温泉"		"バス"		"炭酸鉄泉"
hinagu-spa-j.html	"日奈久温泉"	"JR八代駅"	"JR日奈久駅下車"		"食塩泉" "単純"
kanaketa-spa-j.html	"金鉢温泉"	"JR三角駅"	"バス"		"炭酸鉄泉"
tsurugiyama-spa-j.html	"鶴木山温泉"	"JR佐敷駅"			"単純"
tsuruyu-spa-j.html	"鶴湯温泉"		"徒歩"		"単純"
yoshio-spa-j.html	"吉尾温泉"	"JR吉尾駅"	"徒歩"		"単純"
yunoko-spa-j.html	"湯の子温泉"	"JR水俣駅"	"バス"	"沖合いの湯"	"重曹泉"
yunotsuru-spa-j.html	"鶴温泉"	"JR水俣駅"	"バス"		"単純"
yunoura-spa-j.html	"湯浦温泉"	"JR湯浦駅"	"徒歩"		"単純"

Figure 2: An Example of Reorganization of Hot-Spring Information on the WWW

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 *et al.* 1994), and technical books (Nishiki *et al.* 1994). Figure 3 shows a part of of an ontology about artificial intelligence.

Weakly structured ontologies have only one type of associative relation between terms. Conceptual rela-

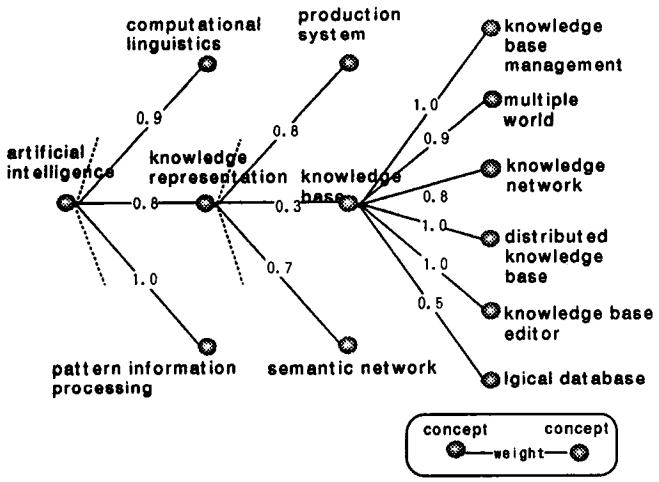

Figure 3: An Example of a Weakly Structured Ontology

tions 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.

Ontology-based intelligent information gathering

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

Information gathering on the WWW

IICA collects WWW 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.

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

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

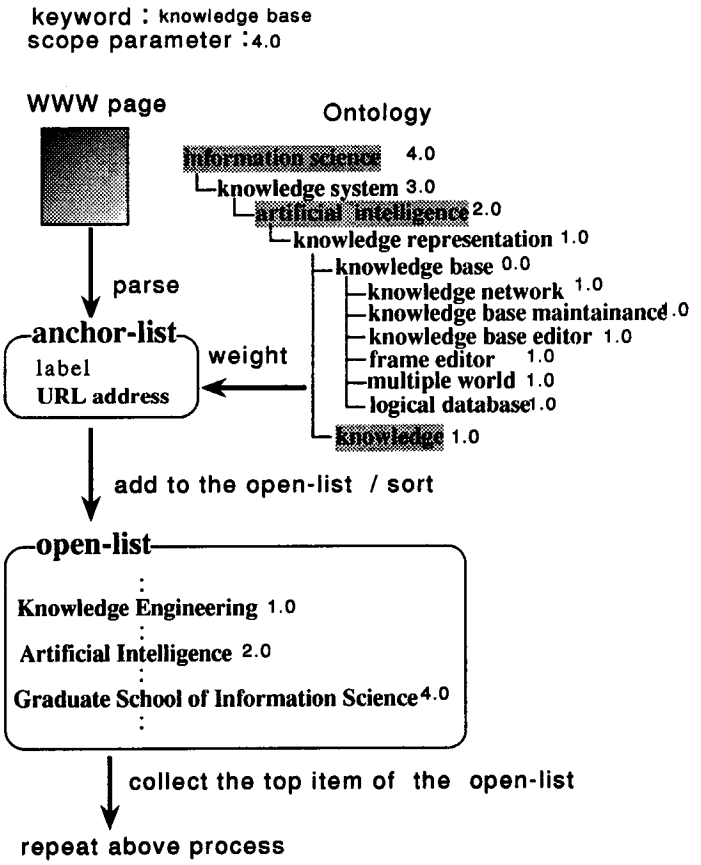

Figure 4: An Example of Information Gathering on the WWW

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.

Figure 4 shows an example of gathering pages on the WWW using the ontology-based method.

Example Suppose that the user's query consists of a keyword "knowledge base"and a scope parameter 4.0. IICA generate a set of related terms to the keyword using the ontology (See the upper right-hand side in the Figure 4). The distance between each related terms and the query keyword is within 4.0. In this example, The anchor "Knowledge Engineering' is given a weight 1.0 because it contains the pattern "knowledge". For detailed technical information, see (Iwazume *et al.* 1995).

Heuristics

When we search for information on the WWW, we use various heuristics such as empirical knowledge and common-sense. For example, the following heuristics seems reasonable when we search information on artificial intelligence.

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

Such kind of heuristics can make the information gathering process more effective in cooperation with ontologies. For instance, the heuristics 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 heuristics.

Ontology-based text categorization

Ontology-based text categorization is the classification of documents by using ontologies as category definition.

In our approach, the process of text categorization is twofold: (1) Text categorization by calculating similarity between a feature vector and a category vector, (2) Modifying weights between terms in a ontology by calculating similarity between category vectors (see Figure 5).

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

We use vector space model commonly used in the information retrieval studies to weight terms and calculate feature vectors (Salton and McGill 1983). The algorithm is as follows:

step1: Calculate the feature vectors of the gathered pages.

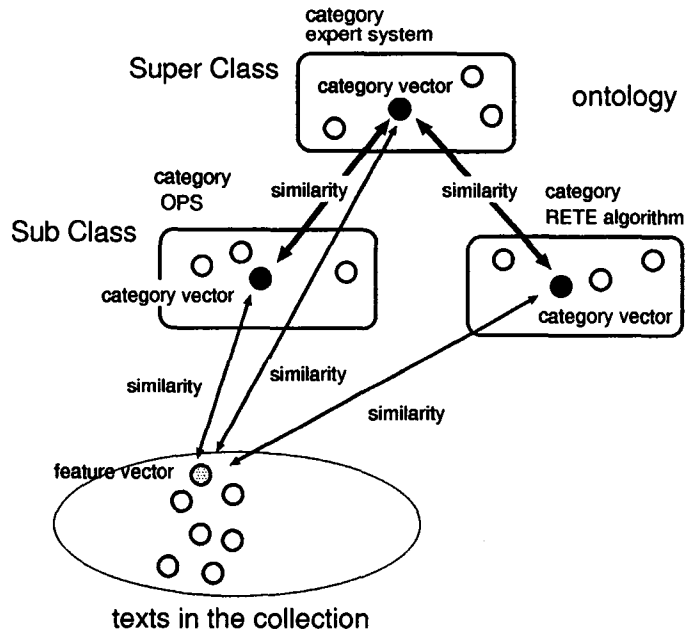

Figure 5: Text Categorization Using an Ontology

step2: Classify the gathered pages by calculated the feature vector.

step3: Calculate the category vectors from the classified pages.

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 vector is calculated from the feature vector of the pages which is assigned to the category by matching keywords.

Information Extracting and Reorganization

This section describes information extracting and reorganization using heuristics. We collected and analyzed the sightseeing pages in Japanese. As the result, it was found that it is possible to extract and reorganize specific information form pages using heuristics based on expression patterns and phrases.

1. State Diagram Method It is the method to analyze and extract specific items according to a state diagram. For example, in case of extracting information about transport facilities, IICA analyzes in such sequence as,
   bus stop(point) → *bus* → *bus stop(point)* → *walk* → ....
2. Rule-base method It is the method to extract specific items according to attributes and rules defined

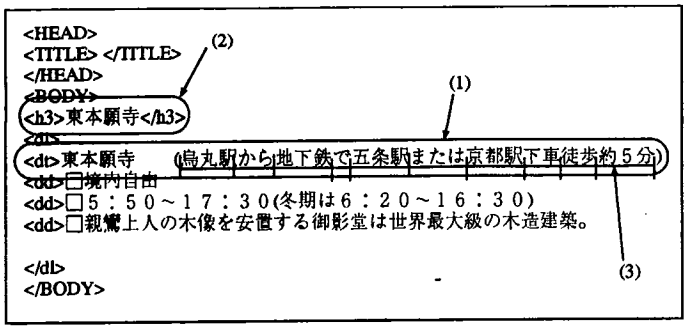

Figure 6: An Example of Extraction of Traffic Information Using A State Diagram

in ontologies. This method can be widely applied to various information on the WWW.

We describe the above two methods in detail.

b

State Diagram Method

The process of this method consists of three steps: (1) finding description, (2) extracting names of sightseeing places, and (3) analyzing description and extracting items using a diagram (See Figure 6).

Figure 7 shows the process of analyzing description about transport facilities. The above state diagram in the Figure 7 is used for analysis and the bottom sentence is the target description. The thick curved line shows a sequence of states in the analysis. The analysis starts at the initial state.

The pattern "駅 (station)"turns out in the description, the current state changes the state 「地点 (point)」 Since the first segment of the description includes "駅 (station)" which indicates "point", the current state changes to the "point" and the system gets the station name "河原町駅 (the Kawaramachi Station)".

Next, the pattern " バス (bus)" is found , the current state changes to the state 「バス (bus)」 and it gets the name of the bus company "市バス (the City Bus)".

Then, since the expression pattern "停 (bus stop)" is found in the description. Therefore the current state changes to the state 「地点 (point)」 and it gets the bus stop name "修学院離宮道停 (the Shugakuin Detached Palace Street Bus Stop)". It repeats the same process till the analysis reaches the end of the description.

Rule-based method

The descriptions such as 「効能は神経痛である (It takes effect on neuralgia)」, 「露天風呂がある (There is an open-air bath)」 appear frequently in the pages about hot-springs. The expression "痛" means a pain and the expression "風呂"means a bath in Japanese. Then we

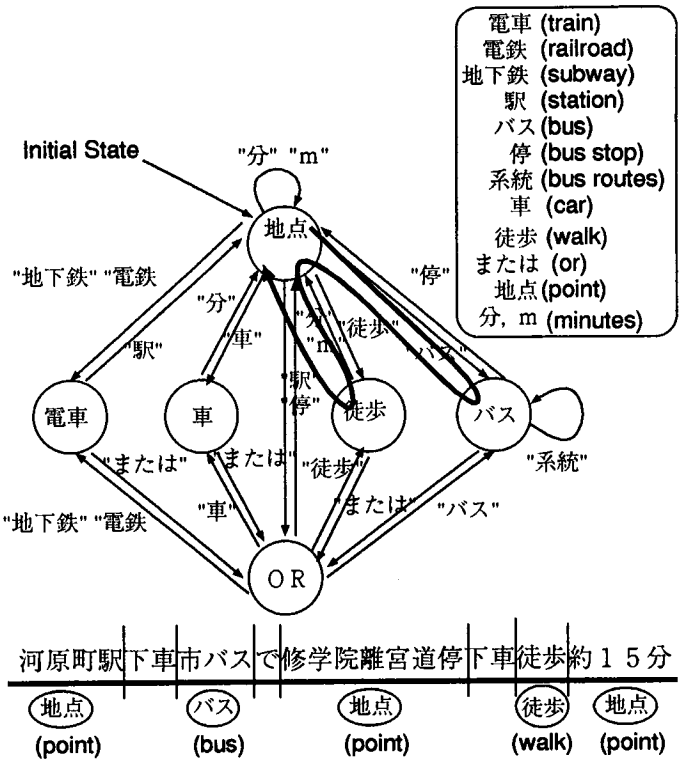

Figure 7: The Process of State Diagram Method

```
(define-pclass
  (温泉
    ((has-one 温泉の名前)
     (is-a 訪問地)
     (has-some 風呂の種類)
     (has-some 泉質)
     (has-some 効能))))
```

Figure 8: A Definition of Extraction Items

use rules based on expression patterns and phrases like the above examples.

The process of describing extraction rules is twofold.

1. Definition of attributes

   A specific item to extract is defined as an attribute of a class in the ontology. For instance, attributes such as name, style, ingredient, effects are defined to extract information related with hot-spring.

   Figure 8 shows the definition of hot-spring attributes. This formula means that 温泉 (hot-spring) has the attribute called 温泉の名前 (name) which take one value, the attributes such as 風呂の種類 (style), 泉質 (ingredient) and 効能 (effects) which take some values, and it is a 訪問地 (a tourist resort).

2. Describing extraction rules based on specific expres-

```
(define-concept
  (効能
    (is 傷病 with
      (or "効能>" "効果" "効く ")))
(define-concept
  (傷病
    (or "+ 症>" "+ 傷>" "+ 病>")))
```

Figure 9: Attribute Extraction Rules for Hot-Springs

Table 1: Evaluation of Gathering Pages Relevant to Artificial Intelligence

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

sion patterns

Figure 9 shows the rules to extract effects of hot-springs. The first rule means that if the expression pattern "効能 (effect)" or "効く (effective)" appears with the concept 傷病 (sickness and injury) in the same sentence, the pattern matching the concept 傷病 indicates 効能 (effects). The second rule means that the expressions patterns "+ 症"or"+ 傷"or"+ 痛" turns out in the sentence, the pattern indicates the concept 傷病 (sickness and injury). Here, the symbol "+" holds the same meaning regular expression. For example, the expression "+ 痛 (pain)' matches "関節痛 (arthralgia)", "腰痛 (lumbago)" and so on.

Evaluation

This section describes evaluation of our method.

Evaluation of Gathering Information

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

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 doesn't use ontologies. It traces hyperlinks on the WWW using breadth first algorithm.
2. Ontology based search: IICA uses ontology-based search algorithm.
3. Ontology based search with heuristics: IICA uses ontology-based search algorithm and heuristics.

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

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 + 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 + heuristics	44	13	25

○: 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.

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 numbers of pages.

Evaluation of Information Categorizing

We 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 5.

Evaluation of Extracting Information

The evaluation of two extracting methods is done with The targets were the WWW pages about sightseeing in Japanese. we tested our state diagram method for analyzing the 100 pages which contained description about transport facilities. Table 6 shows the results of the experiment.

We tested rule-based method for extracting information form the pages concerned with hot-spring, restaurant, and temples. Figure 7 shows recall and precision results.

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 + heuristics	18	23	15

Table 5: Evaluation of Categorization of WWW pages

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

Conclusion

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

We have 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.

We have tested our approach for tasks on the WWW We can conclude the following advantages of our approach from the results.

- Ontology and heuristics make accuracy and efficiency better in information gathering.
- IICA 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 possible to easily extract and reorganize specific information from very large text data by using heuristics based on expression patterns and phrases.
- It is easier to develop weakly structured ontologies from terminologies and thesauruses than conventional methods.
- The ontology-based approach enables us intensive use of heterogeneous information resources on the wide-area such as the WWW.

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

Table 6: Evaluation of Extraction of Traffic Information Using A State Diagram

1. a rate of pages which contain descriptions accurately found	85 %
2. a rate of pages which contain descriptions accurately analyzed and extracted	70 %

Table 7: Recall and Precision of Extraction of Information Using Heuristics

Domain	Precision	Recall
hot-springs	82.2 %	61.2 %
temples	72.2 %	73.4 %
restaurants	85.0 %	41.0 %
Average	79.8 %	58.6 %

References

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

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

Iwazume, Michiaki; Takeda, Hideaki; and Nishida, Toyoaki 1994. Automatic classification of articles in network news and visualization of discussions — intelligent news reader. *Proceedings of the 8th Annual Conference of JSAI* 186–193. (in Japanese).

Iwazume, Michiaki; Takeda, Hieaki; and Nishida, Toyoaki 1995. Ontology-based approach to information gathering and text categorization. *Proceedings of International Synposium on Digital Libraries 1995* 186–193.

McBryan, O. 1994. Genvl and wwww:tools for taming the web. In *Proceedings of 1st International WWW Conference.*

Nishiki, Masanobu; Takeda, Hideaki; and Nishida, Toyoaki 1994. Extraction, unification and presentation ok knowledge by multi agent system. *Proceedings of the 8th Annual Conference of JSAI* 505–508. (in Japanese).

Pinkerton, B. 1994. Finding what people want: Experiences the webcrawler. In *Proceedings of 2nd International WWW Conference.*

Salton, G and McGill, M. J. 1983. Intoroduction to modern infomation retrieval. *MacGraw-Hill.*