IICA: An Ontology-based Internet Navigation System

Michiaki Iwazume, Kengo Shirakami, Hatadani Kazuaki
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: +81-07437-9-5265
Fax: +81-07437-2-5269
E-mail: {mitiak-i, kengo-s, kazuak-h, takeda and nishida }@is.aist-nara.ac.jp

Abstract

In this paper, we propose a new method to develop more intelligent Internet navigation systems using *Ontology*. 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, categorizing and reorganizing information.

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

Keywords: Ontology, the World Wide Web, Information gathering, Text categorization, Reorganization.

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. 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, classifying and 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.

This system (1)Information Gathering: 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

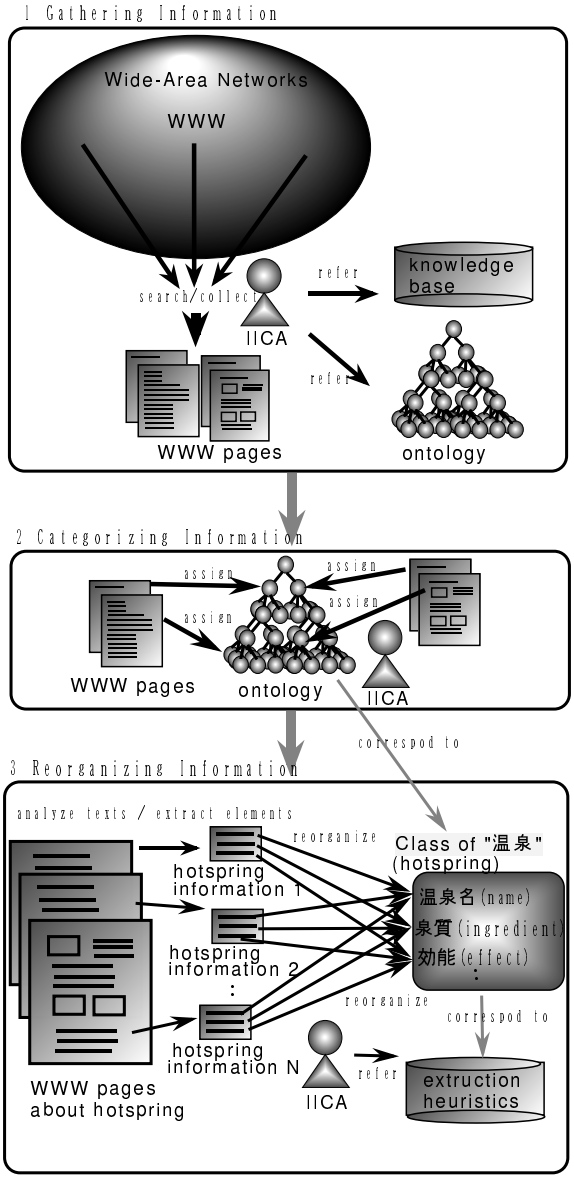

Figure 1: Outline ofIICA

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	"金桁温泉"	"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	"児温泉"	"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

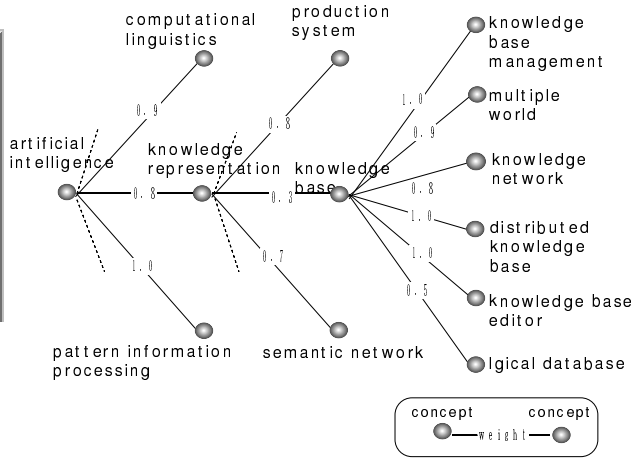

Figure 3: An Example of a Weakly Structured Ontology

texts, (2)Information Categorizing: categorizes the gathered texts by linking them with an ontology and (3)Information Reorganizing: extracts specific information form texts using heuristics based on expression patterns and phrases(See 2).

We tested IICA for tasks on the WWW. The results of the experiments indicated that the ontology-based approach enable us to use heterogeneous information resources on the wide-area networks such as the Internet.

In Section 2, We will explain an information gathering method using ontologies and heuristics. In Section 3, we will explain a new method of text categorization using ontologies. In Section 4, we will describe how IICA uses heuristics based on expression patterns and phrases to extract and reorganization specific information from texts. In Section 5, we will describe the evaluation of the above three methods. In Section 6, we will discuss the advantages of our approach and summarize this paper.

2 Ontology-based intelligent information gathering

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

2.1 Ontology

An Ontology is specification of conceptualization which consists of a vocabulary and a theory [2].

The role of ontologies on information gathering is to provide knowledge for agents to infer information which is relevant to user's requests

Ontologies are often described in frame like languages and knowledge representation languages based on first-order predicate logic such as Ontolingua [1]. Unfortunately, development of ontologies is often a quite painstaking and time consuming task.

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.

In the following experiments, we use the ontology built from the information science terminology which has about 4,500 terms.

2.2 Information gathering on the WWW

In this section, we describe the search algorithm 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.

2.2.1 Algorithm

The algorithm is basically breadth-first searching. The difference is that IICA evaluates gathered pages and decides which anchor to access next. Figure 4 shows an example of gathering pages on the www using the ontology-based method.

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 above right in the Figure 4). The distance between each related terms and the query keyword is withing 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 [4].

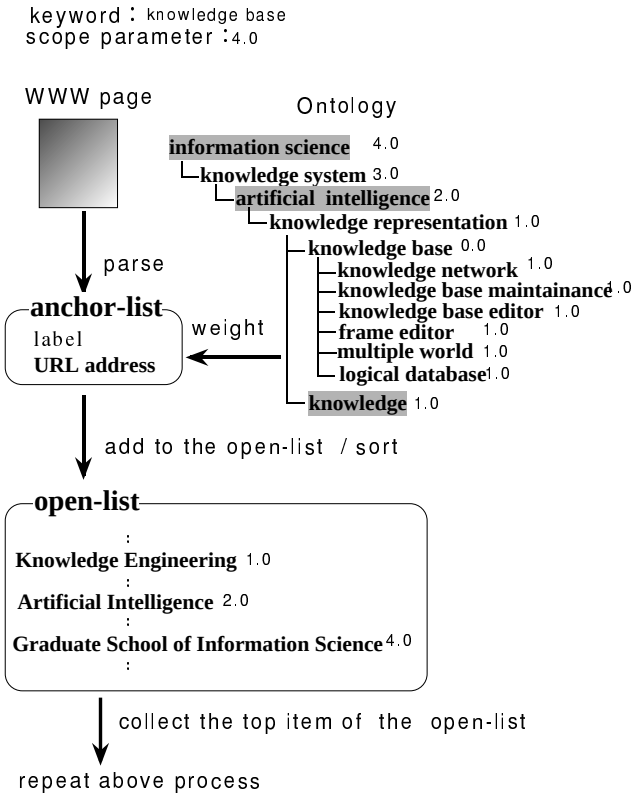

Figure 4: An Example of Information Gathering on the WWW

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 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. 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.

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

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.

4 Information Extracting and Reorganization

This chapter 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 texts 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 analyze 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

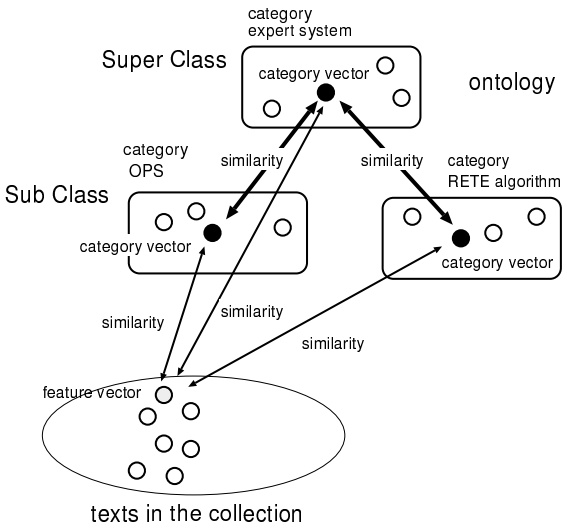

Figure 5: Text Categorization Using an Ontology

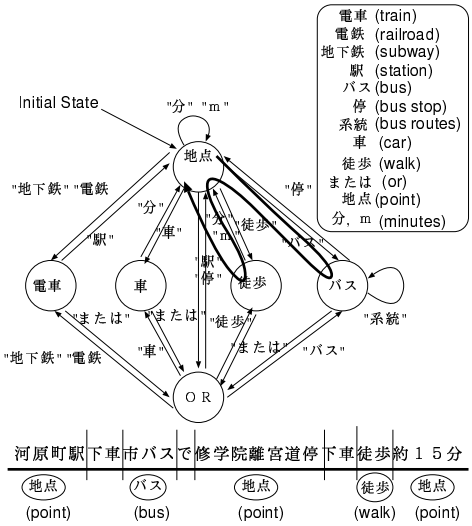

Figure 6: An Example of Extraction of Traffic Information Using A State Diagram

and rules defined in a ontology. This method can be widely applied to various information on the WWW.

We describe the above two methods in detail.

4.1 State Diagram Method

The process of this method is three fold: (1)finding description , (2)extracting names of sightseeing places and (3)analyzing description and extracting items using a diagram.

4.1.1 Analysis and Extraction of Information about Transport facilities

This section explains the process (3). Figure 6 shows the process of analyzing description about transport facilities. The above state diagram in the Figure 6 is used for analysis and the bottom sentence is target description. The bold arrow 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)」 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, the expression pattern "停 (bus stop)" turns up in the description, the current state changes 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 reachs the end of the description.

4.2 Rule-based method

This section explains 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 hotsprings. The expression "痛" means a pain and the expression "風呂"means a bath in Japan. Then we dececided to make use of rules based on expression patterns and phrases like the above examples.

4.2.1 Description

The process of describing extraction rules is twofold.

1. Definition of attributes

   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 hotspring.

   Figure 7 shows the definition of hotspring attributes. This formula means that 温泉 (hotspring) 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 expression patterns

   Figure 8 shows the rules to extract effects of hotsprings. The first rule means that if the expression pattern "効能 (effect)" or "効く (effective)" appears with the concept

```
(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 7: A Definition of Extraction Items

傷病 (sickness and injury) in the same sentence, the pattern matching the conept 傷病 indicates 効能 (effect). 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.

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

Figure 8: Attribute Extraction Rules for Hot-Springs

5 Evaluation

This chapter describes evaluation of our method.

5.1 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.

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 intrests.

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 + 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 + 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

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 Information Categorizing

We made an experiment of categorizing the about 500 pages concerned with AI in English

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

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.

Table 5: Evaluation of Categorization of WWW pages

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

5.3 Evaluation of Extracting Information

This section describes the evaluation of two extracting methods. The targets were the WWW pages about sightseeing in Japanese.

5.3.1 Evaluation of State Diagram Method

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.

5.3.2 Evaluation of rule-baded method

We tested rule-based method for extracting information form the pages concerned with hotspring, restaurant, temples. Figure 7 shows Recall and Precision.

6 Conclusion

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

Table 6: Evaluation of Exatraction 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 analysed and extracted	70 %

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

Domain	Precision	Recall
hotsprings	82.2 %	61.2 %
temples	72.2 %	73.4 %
restaurants	85.0 %	41.0 %
Average	79.8 %	58.6 %

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.

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

- Ontology and heuristics make accurcy 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 possible to easily extract and reorganization specific information from very large text data by using heuristics based on exression patterns and phrases.
- 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 wide-area such as the WWW.

The problem of the current system is that ontologies 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

[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] Michiaki Iwazume, Hieaki Takeda, and Toyoaki Nishida. Ontology-based approach to information gathering and text categorization. *Proc. of ISDL95*, 1995.

[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.