Ontology-based Information Capturing from the Internet

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.

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

2. Ontology-based Intelligent Information Gathering

2.1 Ontology

An Ontology is specification of conceptualization which consists of a vocabulary and a theory (Gruber, 1991). 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 (Gruber, 1992). We decided to make use of weakly structured ontologies (Nishida, 1995) which is developed from existing terminologies, thesauruses (Iwazume, 1994). Weakly structured ontologies have only one type of associative relation between terms (see Figure 3).

2.2 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. 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 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 (1995).

2.3 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.

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, 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:

Step 1. Calculate the feature vectors of the gathered pages.
Step 2. Classify the gathered pages by calculated the feature vector.
Step 3. Calculate the category vectors from the classified pages.
Step 4. Repeat step 2 and step 3 until the category vectors converge.
Step 5. Calculate distance between the categories and renew weight between terms in the ontology.

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

4. 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 from pages using heuristics based on expression patterns and phrases.

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 → ...

Rule-based method. It is the method to extract specific items according to attributes and rules defined in ontologies. 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 consists of three steps: (1) finding description, (2) extracting names of sightseeing places, and (3) analyzing description and extracting items using a diagram.

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 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 to 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 " 情 " (bus stop)" is found in the description, 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.

4.2 Rule-based Method

The descriptions such as 効能は神経痛である ("It takes effect on neuralgia") and 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 use rules based on expression patterns and phrases like the above examples.

The process of describing extraction rules is twofold.

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.

The following formula is the definition of hot-spring attributes. This formula means that 溫泉 ('hot-spring') has the attribute called 溫泉の名前 ('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 ('tourist resort').

\[
\text{(define-pclass (温泉 ((has-one 温泉の名前 )}} \\
\text{(is-a 訪問地) )}} \\
\text{(has-some 風呂の種類 )}} \\
\text{(has-some 泉質 )}} \\
\text{(has-some 効能 ))))}
\]

Describing extraction rules based on specific expression patterns. The following formulas are 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 " + 病 " 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.

\[
\text{(define-concept (効能 (is 傷病 with (or " 効能 >" " 効果 " " 効く )))}} \\
\text{(define-concept (傷病 (or " + 症 >" " + 病 >" " + 痛 >")))}
\]

5. Evaluation

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 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.

- $\bigcirc$: The collected page is directly related to user's queries.
- $\triangle$: The collected page is not directly related to user's queries, but it is related to user's interests.
- $\times$: The collected page is neither directly related to the user's queries nor related to user's interests.

Table 1 and Table 2 show the results.

<table>
<thead>
<tr>
<th>Search method</th>
<th>$\bigcirc$ (%)</th>
<th>$\triangle$ (%)</th>
<th>$\times$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Breadth first search</td>
<td>64.6</td>
<td>07.4</td>
<td>28.0</td>
</tr>
<tr>
<td>2. Ontology</td>
<td>66.6</td>
<td>11.6</td>
<td>21.8</td>
</tr>
<tr>
<td>3. Ontology + heuristics</td>
<td>67.8</td>
<td>10.6</td>
<td>21.6</td>
</tr>
</tbody>
</table>

Table 1: Evaluation of Gathering Pages Relevant to Artificial Intelligence

<table>
<thead>
<tr>
<th>Search method</th>
<th>$\bigcirc$ (%)</th>
<th>$\triangle$ (%)</th>
<th>$\times$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Breadth first search</td>
<td>57.4</td>
<td>08.4</td>
<td>34.2</td>
</tr>
<tr>
<td>2. Ontology</td>
<td>59.5</td>
<td>15.6</td>
<td>24.9</td>
</tr>
<tr>
<td>3. Ontology + heuristics</td>
<td>59.5</td>
<td>15.6</td>
<td>24.9</td>
</tr>
</tbody>
</table>

Table 2: Evaluation of Gathering Pages Relevant to Sightseeing

### 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 numbers of pages.

### 5.2 Evaluation of Information Categorizing

We made an experiment of categorizing the about 500 pages concerned with AI in English and the about 800 pages concerned with sightseeing in Japanese. In order to evaluate our method, we calculated recall and precision. The result is shown in Table 5.
Table 3: Evaluation of Efficiency of Information Gathering—
1 keyword ("knowledge base")

<table>
<thead>
<tr>
<th>Search method</th>
<th>○</th>
<th>Δ</th>
<th>×</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Breadth first search</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2. Ontology</td>
<td>21</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>3. Ontology + heuristics</td>
<td>44</td>
<td>13</td>
<td>25</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Search method</th>
<th>○</th>
<th>Δ</th>
<th>×</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Breadth first search</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2. Ontology</td>
<td>10</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>3. Ontology + heuristics</td>
<td>18</td>
<td>23</td>
<td>15</td>
</tr>
</tbody>
</table>

5.3 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 also tested rule-based method for extracting information from the pages concerned with hot-spring, restaurant, and temples. Table 7 shows recall and precision results.

6. Conclusion

In this paper, we present a system called IICA which gathers, classifies, and reorganizes information from the Internet. We have implemented and evaluated IICA. 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 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.
- The ontology-based approach enables us intensive use of heterogeneous information resources on the Internet such as the WWW.
### Table 5: Evaluation of Categorization of WWW Pages

<table>
<thead>
<tr>
<th></th>
<th>AI (English)</th>
<th>Sightseeing (Japanese)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>81.9</td>
<td>79.0</td>
</tr>
<tr>
<td>Recall</td>
<td>80.5</td>
<td>70.0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Domain</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot-springs</td>
<td>82.2%</td>
<td>61.2%</td>
</tr>
<tr>
<td>temples</td>
<td>72.2%</td>
<td>73.4%</td>
</tr>
<tr>
<td>restaurants</td>
<td>85.0%</td>
<td>41.0%</td>
</tr>
<tr>
<td>Average</td>
<td>79.8%</td>
<td>58.6%</td>
</tr>
</tbody>
</table>

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

**References**


Figure 1: Outline of IICA

Figure 2: An Example of Reorganization of Hot-Spring Information on the WWW
Figure 3: An Example of a Weakly Structured Ontology
keyword: knowledge base
scope parameter: 4.0

WWW page

Ontology

- information science 4.0
- knowledge system 3.0
- artificial intelligence 2.0

knowledge representation 1.0

- knowledge base 0.0
- knowledge network 1.0
- knowledge base maintenance 0
- knowledge base editor 1.0
- frame editor 1.0
- multiple world 1.0
- logical database 1.0

knowledge 1.0

parse

anchor-list

label

URL address

weight

add to the open-list / sort

open-list

Knowledge Engineering 1.0
Artificial Intelligence 2.0
Graduate School of Information Science 4.0

collect the top item of the open-list

repeat above process

Figure 4: An Example of Information Gathering on the WWW
Figure 5: Text Categorization Using an Ontology
Figure 6: An Example of Extraction of Traffic Information Using A State Diagram