Improving Rule Generation Precision for Domain Knowledge based Wrappers

Chang-Hoo Jeong*, Sung-Jin Jhun*, Myung-Eun Lim**, Sung Hyon Myaeng***
Korea Institute of Science and Technology Information (KISTI),*
*Electronics and Telecommunications Research Institute (ETRI)**,*
*Information and Communications University (ICU)****
{chjeong, sjjhun}@kisti.re.kr, melim@etri.re.kr, myaeng@icu.ac.kr

Abstract

Wrappers play an important role in extracting specified information from various sources. Wrapper rules by which information is extracted are often created from the domain-specific knowledge. Domain-specific knowledge helps recognizing the meaning the text representing various entities and values and detecting their formats. However, such domain knowledge becomes powerless when value-representing data are not labeled with appropriate textual descriptions or there is nothing but a hyper link when certain text labels or values are expected. In order to alleviate these problems, we propose a probabilistic method for recognizing the entity type, i.e. generating wrapper rules, when there is no label associated with value-representing text. In addition, we have devised a method for using the information reachable by following hyperlinks when textual data are not immediately available on the target web page. Our experimental work shows that the proposed methods help increasing precision of the resulting wrapper, particularly extracting the title information, the most important entity on a web page. The proposed methods can be useful in making a more efficient and correct information extraction system for various sources of information without user intervention.

1. Introduction

The amount of information on the Internet has increased dramatically with the improvement of Internet technology. Anyone, even without sophisticated knowledge on computers, can own a web site and web service providers are providing various types of information in all areas. With the increase in the amount of data, it has become difficult for the user to find the core information they are looking for. Such overload of information decreases user-satisfaction on Internet service and makes it difficult to retrieve the required data even using existing search engines. The need for a new system that extracts the required information to provide to the user is being raised.

The most general way of accessing information on the Internet is through creating a capsulized wrapper that can access the various sources of information. Wrapper can be defined as the rule for extracting the location, structure and format of the data. The existing domain knowledge based wrapper learning method creates the wrapper using the information on the domain. Domain knowledge defines the entity, which is a concept used for a specific application domain, and selects the label and describes the format for the data. The system can acquire the meaning of the texts provided from the source of information as well as the structure by using the domain knowledge defined for each area of application. However, since the label is not provided for all the texts provided from the source of information, the method using the domain knowledge is limited.

2. Related Studies

Researches on wrapper generation include studies on manual creation, semi-automatic creation and automatic creation methods[1]. In the manual method, a human describes the rules for the extraction[2], and the semi-automatic method creates the wrapper using the tools used for the design[3]. The automatic method uses the learning method for automatically creating the wrapper. Various algorithms have been developed for the automatic creation of wrapper, which are time consuming since they have to perform learning, but can minimize the effort put in by human experts.

Automatic creation methods include the method that classifies the classes that can be extracted before the learning[4] and the domain-knowledge based method[5]. The first method adopts wrapper induction and defines a few wrapper classes before processing the source of information. There are 6 classified classes for which a wrapper generation algorithm, learn-W, is provided. The domain knowledge based method creates a domain knowledge for each domain and suggests a method for creating the wrapper for each of them. Since the information to be extracted are already expressed in the domain knowledge, there is no need to construct the learning data. However, the documents from the source of information have to have a label in order to use the domain knowledge.

In this paper, we will describe the domain knowledge based probabilistic wrapper generation system used to effectively extract the information from a semi-structured source of information on the web. We will suggest a probabilistic model that automatically recognizes the entity of the text without a label. This method is similar to the domain knowledge based method in that it minimizes human effort which makes it easy to apply to real-world situations. It also performs the recognition of entities without a label that the domain knowledge based method could not perform.

3. Wrapper Generation

The factors to consider in order to improve the performance of the wrapper generation are as follows.

3.1. Method using the Hyperlink

Most sources of information on the web only provide the abstract at first. The specific information can be viewed through a hyperlink. This method enables the user to quickly look through the information that the user demanded. In order to provide the user with as much information as possible on the initial page, the information should be obtained from the database to create the web page which is a time-consuming event. This would cause more harm than good to the user due to the long initial access time. Therefore, the information linked to the hyperlink should be utilized effectively in order to acquire the necessary information.

The methods for using the hyperlink are as follows.

- When creating the wrapper
  - Detect the boundary of the item by analyzing the pattern of the information provided in the main page.
  - Follow all the hyperlinks within the detected boundary to check for useful information. A useful information is a document with the largest number of entities acquired using the domain knowledge.
  - If the information is found to be useful, the location of the link and the information related to the entities are merged to be recorded on the wrapper.
- When extracting the information
  - Read the wrapper information in order to decide whether to extract information through the hyperlink.
  - Extract the information from the main page, and acquire further information from the page connected through the hyperlink if there is a sign.
  - Combine the information from the front page and the back end page to create a single item of information extracted.

The number of useful entities can be increased by using the analyzed information included in the hyperlink.

3.2. Probability Information Utilization Method

Text with labels is automatically recognized by the domain knowledge in creating a wrapper from the source of information. However, those without a label cannot be used for recognizing the entity since they do not contain any clue to their meaning. In this section, we introduce a probabilistic method to recognize the meaning of texts that cannot be detected using the above method.

3.2.1. Background. Text on a page can be distinguished by the possession of a label. The meaning and structure information of a text with a label can be automatically recognized. Those without a label can adopt the probabilistic method in order to acquire the information.

We will first define the term used to describe the model. An entity is the basic unit of the information used in the domain. A label is a clue to the recognition of the entity and an item is the basic unit of information provided by the source of information. Usually, most sources of information display the contents on the web page with a certain pattern such as list or table. An item can also be defined as the tuple of the database. The pieces of text appear separately through tags when the structure analysis is applied. These pieces are logically restructured to form a text with a meaning as shown on the browser. The part which can become the value of the entity is called the token. A number of tokens are created by performing a structural analysis of the HTML document. In this step, both texts with and without the label information appear in a same pattern to form a set of tokens for the source of information. When a token is selected for a single item, a token at the same location of another item can be defined as performing the same role. These are called a token set. Therefore, a number of token sets can be found within a single source of information. It is called a token set sequence since a number of token sets appear sequentially. The process for forming a token set sequence is shown in Figure 1 below.

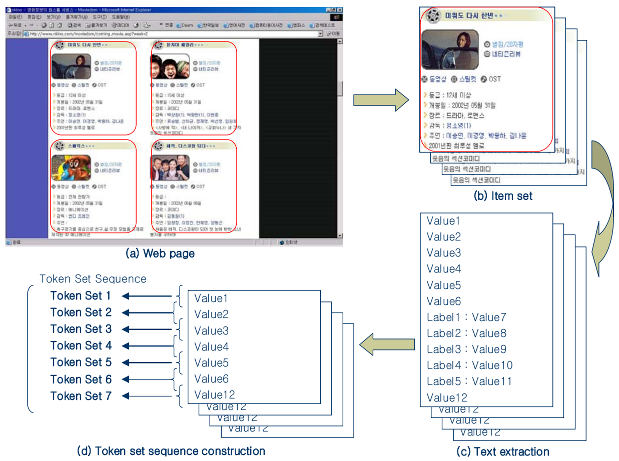

Figure 1. Step for creating the token set sequence

(a) of Figure 1 is the result of browsing the HTML document which is a pattern easily found in the source of web information. (b) shows the result of re-arranging the information of interest by item. More than one item is acquired from a single source of information. (c) shows the result after grouping the text information by extracting a certain pattern through analyzing the structure of the web page. When performing the structural analysis of the document in order to create such information, the text information has to be linked logically as shown in (b). Link tags and font tags between texts have to be removed in order to extract the text. (d) shows the token set from the item which can become the entity value. Since there is more than one token set, it shows that a token set sequence can also be formed.

Figure 2 shows the result of expressing the item on Figure 1 as a tuple of a database.

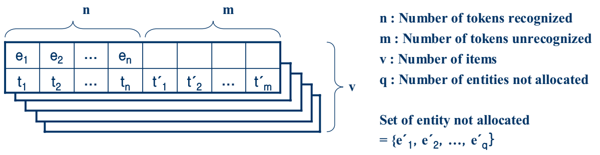
n
m
e1
e2
…
en
t1
t2
…
tn
t´1
t´2
…
t´m
v
n : Number of tokens recognized
m : Number of tokens unrecognized
v : Number of items
q : Number of entities not allocated
Set of entity not allocated
= {e´1, e´2, …, e´q}

Figure 2. Tuple expression of items

The number of tokens not recognized for each item is expressed as m, as shown in Figure 2. The main point of this model is on classifying the unrecognized tokens to each set of entity.

The rules can be defined as follows.

1. There are n recognized tokens for a single item. $\{t_1, t_2, \ldots, t_n\}$

2. There are n allocated entities. $\{e_1, e_2, \ldots, e_n\}$

3. The number of unrecognized token for a single item is m. $\{t'_1, t'_2, \ldots, t'_m\}$

4. The number of entities not allocated is q. $\{e'_1, e'_2, \ldots, e'_q\}$

e'_k is the set of entities formed by removing the entities found in the source of information from the set of entities E. The entity of the token is given exclusively, which means that the entities already found have to be removed from the set of entities that can be newly recognized.

5. There are v items in a single source of information.

6. There are n recognized token sets for a single source of information. There are v tokens in a token set. $\{T_1, T_2, \ldots, T_n\}$, $T_i = \{t_{i1}, t_{i2}, \ldots, t_{iv}\}$

7. There are m unrecognized token sets for a single source of information. There are v tokens in a token set. $\{T'_1, T'_2, \ldots, T'_m\}$, $T_j = \{t'_{j1}, t'_{j2}, \ldots, t'_{jv}\}$

8. A set E with (n + q) number of entities is defined in the domain knowledge.

We will suggest a probability model based on the above assumptions.

3.2.2. Designing the Entity Recognition Model.

The ERM(Entity Recognition Model) suggested in this study was inspired by the HMM(Hidden Markov Model). HMM performs the tagging of category on each word in a sentence[6]. The ERM suggested in this study provides an entity to each token that forms a single item. The difference is that it does not apply the statistical method on every word. The tokens with a label are excluded in applying the statistical method. The other main difference is that the HMM takes the order of the elements into account whereas the ERM does not take order of the tokens into account. Therefore, it does not require the minimization of the number of possible instances through the Viterbi algorithm. Due to these differences, the equation of the ERM is different from that of HMM. The difference between the ERM and HMM is shown in Table 1.

Table 1. Comparison of HMM and ERM

	HMM	ERM
Class	Category	Entity
Object	Word	Token
Probability	Probability that category is tagged onto word	Probability that entity is recognized onto token
Corpus Statistics	Lexical Generation Probability	Model 1 Probability
Context Information	Bigram Probability	Model 2 Probability
Difference	Occurrence order of category is important.	Occurrence order of entity isn't important.

$$
\begin{aligned}
&HMM \\
&= PROB(C_1,\ldots,C_T \mid w_1,\ldots,w_T) \\
&\cong \prod_{i=1,T} PROB(w_i \mid C_i) * PROB(C_i \mid C_{i-1})
\end{aligned}
$$

$$
\begin{aligned}
& ERM \\
& = PROB(e'_1,\dots,e'_q \mid T'_1,\dots,T'_m) \\
& \cong \alpha * \{P(e'_i) * \frac{1}{v}\sum_{k=1}^{v} P(t'_{jk} \mid e'_i)\} + \\
& (1-\alpha) * \{\frac{1}{v}\sum_{k=1}^{v}\sum_{h=1}^{n} P(e'_i = t'_{jk} \mid e_h = t_{hk}) * P(e_h = t_{hk})\}
\end{aligned}
$$

(1 <= i <= q and 1 <= j <= m)

The probability that part-of-speech is tagged onto word in the HMM is expressed using the Lexical Generation Probability: $PROB(w_i \mid C_i)$. The same is expressed using the Model 1 Probability: $P(e'_i) * \frac{1}{v}\sum_{k=1}^{v} P(t'_{jk} \mid e'_i)$ in the ERM. The Bayesian model uses the conditional probability to classify the entities of a token without a label based on the history of the previous tokens. However, since there are more than just one item on the web page, all the tokens on the same location have to be taken into account. Acquiring the possibility of a set of tokens with the same characteristics would be more appropriate for an accurate result. We would be able to allocate the tokens without a label to a new entity using the concept explained above.

The Bigram Probability: $PROB(C_i \mid C_{i-1})$ used by the HMM to show the probability that two categories appearing sequentially can be expressed using the Model 2 Probability: $\frac{1}{v}\sum_{k=1}^{v}\sum_{h=1}^{n} P(e'_i = t'_{jk} \mid e_h = t_{hk}) * P(e_h = t_{hk})$ in the ERM. This method uses the information of a single item to classify a token without a label by using the text information of a token with a label within the same item. This is possible because the label of an unrecognized token can be assumed by using the information of an already recognized text information. The extracted items contain relative information. The wrapper is created for a various source of information to extract the information to resolve the problems occurring at the current source of information.

As a result, e'_i , which is the largest probability of $PROB(e'_1 ,\dots,e'_q \mid T'_1 ,\dots,T'_m)$, is selected and allocated as an entity of the token set T'_j. However, if the probability is lower than a certain threshold, the token is not recognized at all. The threshold value is used to identify the token as useful or not useful. This threshold value is acquired through experiments.

Finally, the token set T'_j is removed from the first row of tokens to acquire a new token set T'_1, T'_2, ..., T'_{m-1}. The above procedure is repeatedly applied to a newly created token set.

Since ERM is not limited by the order of the elements as in the HMM, the two possibilities are merged together using variable α. Model 1 and Model 2 both have a worthy logic, but a more reliable model with more information can be formed by combining the two. The α value, which shows the relative importance of Model 1 and Model 2 are acquired through experiments.

4. Experiment

The algorithm suggested in this paper has been applied to seven different sources of information in the movie domain (Site A, Site B, …, Site G). The entity of the domain knowledge should be selected appropriately to the application area of the system. Since the purpose of this study is on measuring the performance of the entity recognition model, we have performed the experiments using the largest domain knowledge in the movie area. The entities defined for the movie domain in this study are title, genre, producer, actor, rated level, production, scenario, filming, music, runtime, open date and close date.

4.1. Experiment Method

Experiments have been performed sequentially in order to prove the effectiveness of the methods suggested in this paper. The domain knowledge has been used to create the wrapper in the first stage and then the processing hyperlink was added for the creation of the wrapper. Finally, the entity recognition algorithms have been applied to unrecognized tokens to create the wrapper for the comparison.

The performance of each site can be calculated using the following equation.

※ Precision = (number of extracted entities / number of defined entities in the domain knowledge) * 100

The number of extracted entities is the number of entities recognized during the process of learning by the wrapper, and the number of entities to be extracted is the number of entities defined in the domain knowledge. There was not a single source of information that provided us with all 12 entities and therefore a 100% precision was not possible to achieve.

The average precision of all the sites can be calculated using the following equation.

※ Average Precision = (sum of precision of each site / number of sites)

4.2. Results and Analysis

The result of the precision improvement of each site is shown in Figure 3.

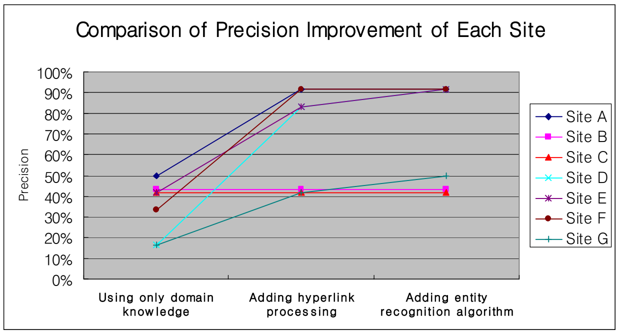

Figure 3. Comparison of precision improvement of each site

The result of the average precision improvement is shown in Figure 4.

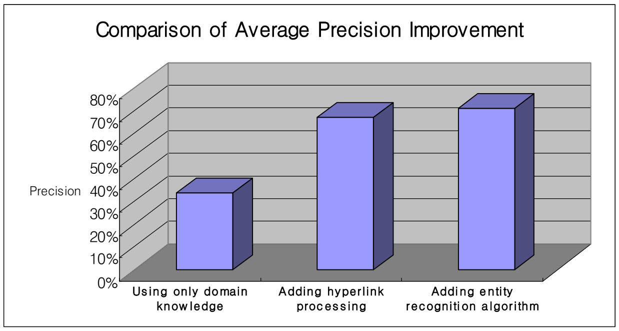

Figure 4. Comparison of average precision improvement

The domain knowledge was applied to create the wrapper in the first experiment. The wrappers were created for the entities extracted from the source of information. However, this method was not able to utilize the hyperlink of the web sites which made it inefficient. Therefore, the number of entities extracted was limited.

The second experiment included the processing of hyperlinks for the creation of the wrapper. The number of entities from some of the sources of information doubled as a result. The structural features of the web site was taken into account through this method, which lead to the improvement of performance. One of the major elements in the improvement of web technology was hyperlink, which was proven in this experiment using the wrapper generation system.

The third experiment applied the entity recognition algorithm for the tokens in order to create the wrapper. The number of entities extracted increased as a result. It proved that applying the probabilistic method on tokens without a label lead to the increase in performance.

The characteristics of the newly recognized entities have to be examined. Information such as the title is a core entity which should exist on any domain, which means that the information would be useless without these entities. However, it was found that the label was missing on important information such as the title in many different sources of information. This was due to the different font or color used to amplify the title in the text. Also, entities used to distinguish the items can be recognized easily by the human eyes which is why the label was missing on such entities. By applying the probabilistic entity recognition method on these documents made it possible to recognize the titles and other important entities.

The other fact found out through the experiments was that the value of Model 1 and Model 2 was different on each source of information. The precision of Model 1 turned out to be larger on some sites whereas the value of Model 2 was larger on other sites. This is assumed to be due to the characteristic of the information provided by each site. In cases where the value of Model 2 was larger than Model 1, the number of texts containing a label in the data was relatively larger. This means that a data with more label contain more context information. The relative importance of Model 1 and Model 2 depends on the characteristic of the data provided by the source of information. Therefore, the variable α is to be applied on the source of data by selecting the model with more importance according to the characteristic of the data.

5. Conclusion and Future Works

The wrapper generation system using the probabilistic method is very important since the method that only used the domain knowledge was proven to be ineffective on text without a label, despite some of its advantages. This is especially true since the unrecognized entities by the system turned out to be of great importance, such as identifiers and titles.

Future works should concentrate on merging the information acquired separately from each source of information and the precision of the rule creation. Information extracted from different sources of information are used to express different items in most cases. However, in some cases, different entities were found for the same item. Therefore the identifier of the extracted item should be used to merge the data for creating the information. Also, new characteristics can be added to the domain knowledge with the change in the source of information and its dynamic characteristics. Such change in the domain should be automatically detected and applied to extend the domain knowledge automatically in future systems.

6. References

[1] L. Eikvil, “Information Extraction from World Wid e Web”, A Survey, 1999.

[2] J. Hammer, H. Garcia-Molina, S. Nestorov, R. Yer neni, M. Breunig, and V. Vassalos, “Template-base d Wrappers in the TSIMMIS System”, ACM SIGM

OD International Conference on Management of Data, 1997, pp. 532-535.

[3] L. Liu, C. Pu, and W. Han, “XWRAP: An XML-enabled Wrapper Construction System for Web Information Sources”, Proceedings of the 16th International Conference on Data Engineering, 2000.

[4] N. Kushmerick, D. Weld, and R. Doorenbos, “Wrapper Induction for information extraction”, International Joint Conference on Artificial Intelligence (IJCAI), Nagoya, Japan, 1997.

[5] H. Seo, J. Yang, and J. Choi, “Knowledge-based Wrapper Generation by Using XML”, IJCAI-2001 Workshop on Adaptive Text Extraction and Mining (ATEM 2001), Seattle, USA, 2001, pp. 1-8.

[6] James Allen, *Natural Language Understanding (2nd Edition)*, Addison-Wesley Publishing Co, 1995, pp. 189-204.