Lexical semantic based Bayesian model for adaptive wrapper generation

Nandhi kesavan.R, Dr.Latha.K
Computer Science and Engineering Department
Anna University of Technology
Tiruchirappalli
nandhikesavan2903@yahoo.com, erklatha@gmail.com

ABSTRACT: - This paper focuses on an unsupervised information extraction system. Two kinds of features related to the text fragments from the Web documents are investigated. The first type of feature is called, a site-invariant feature. The second type of feature is called a site-dependent feature. Feature selection algorithm is used for wrapper generation from the site invariant and site-dependent information. The wrapper is generated and the new attribute is also discovered and adapted with wrapper by using the Bayesian learning method and E-M algorithm along with the lexical semantics search method. Our wrapper can be able to adapt with the new unseen sites. Our system efficiency is evaluated based on some performance measures and the effectiveness of the system is evaluated by using the performance metrics, precision, recall, f-measure, true positive and false positive in the real time web sites.

Key Words:- Unsupervised information extraction, Bayesian learning, E-M algorithm, lexical semantics.

1 INTRODUCTION
An information extraction system aims at automatically extracting precise and exact text fragments from documents. The data on web is highly unstructured and it is not possible with our traditional query engines to query it in an efficient and accurate manner. The available wrapper learning methods are manually-constructed information extraction Systems, supervised information extraction Systems, semi-supervised information extraction Systems [11] and unsupervised information extraction Systems. A major problem with manually-constructed, supervised information extraction Systems and semi-supervised information extraction Systems are required human effort to examine each of the returned entries to extract the precise information. A wrapper [9] [10], normally consists of a set of extraction rules that can identify the text fragments in the documents. It learns a set of extraction rules from the manually annotated training examples. This paper investigates Wrapper adaptation which aims at automatically adapting a previously learned wrapper from one Web site, known as a source Web site, to new unseen sites in the same domain.

![Figure1. A sample Web page of a book catalog](image)

This paper is organized as follows: Section 2 provides existing methods of wrapper generation techniques. Section 3 discusses the new methodology used for developing adaptive wrapper. In Section 4 and 5 we formally evaluate various aspects of our proposed approach based on precision, recall, f-measure, true positive and false positive, and compare the results of our new wrapper generation technique with several other wrapper generation techniques to prove the effectiveness of our wrapper generation technique. Finally in section 6 we conclude that our newly developed technique is more efficient than other techniques and describes direction of the future work.
EXISTING METHODS

2.1 ROADRUNNER
In ROADRUNNER [2] that the generated wrapper extracts data records in an automatic way. In wrapper generation pattern discovery is based on the study of similarities and dissimilarities between the pages. The mismatches are used to identify relevant structures. The system generates a wrapper by examining one HTML page at a time. However, they suffer from a major drawback that they cannot differentiate the type and the meaning of the information extracted. Hence, the items extracted require human effort to interpret the meaning.

2.2 KNOWITALL
KNOWITALL [4], [Etzioni et al., 2005] uses a set of domain-independent extraction patterns to create its set of rules and “discriminator” phrases for each predicate. The two main KNOWITALL modules are the Extractor and the Assessor. The Extractor creates a query from keywords in each rule, sends the query to a Web search engine, and applies the rule to extract information from the resulting Web pages. The Assessor bases its probability computation on search engine hit counts used to compute the mutual information between the extracted instance of a class and a set of automatically generated discriminator phrases associated with that class.

2.3 TEXTRUNNER
The TextRunner [6] system uses the following three-steps for extracting facts. The three key modules that are self-supervised learner, single pass extractor and redundancy based assessor. One common limitation of KNOWITALL and TextRunner is that they cannot be applied to extract attributes and values associated with a particular record from a Web page.

2.4 BAYES WITH E-M METHOD
The unsupervised information extraction [10], system extracts the text fragments from the web sites in an unsupervised manner. It does not require any hand tagged examples. The labels are assigned to the text fragments and it extracts text from web sites based upon the Bayesian learning framework. The E-M algorithm based Bayesian method identifies the new attributes from the new unseen sites [3] the attributes are well refined based on E-step and M step in that algorithm. The limitation is that it cannot distinguish the existing attribute from the semantic meaning of the new attribute from unseen website.

PROPOSED METHOD

In this section we present a new methodology for information extraction. First the method is starting with a clustering method and then picking the sample pages for wrapper generation by using greedy algorithm based upon the Xpath of web pages, And then the preprocessing the pages for labeling the text fragments. The Bayesian learning framework is for generating wrapper and the E-M algorithm for wrapper adaption. In wrapper adaption new attributes are discovered based on E-M algorithm. And the discovered new attributes are adapted to the wrapper by the new attribute discovery approach and lexical semantics method.

3.1 WEB PAGE CLUSTERING
A single Web site may contain pages conforming to multiple different templates. A shingle-based signature [1],[5], is computed for each Web page based on HTML tags in the page, and the pages are clustered using the signatures. For the two pages p and pꞋ the shingle vectors respectively v[i] = v�atformula. The sample web pages are selected based on their XPath. The Web pages are preprocessed before selecting text fragments. Each text fragment can be represented by a three-field tuples (Y, F, C), where Y is the label of the text fragment referring to the attribute to which the contained tokens belong, and F and C refer to the formatting feature and the content feature of the text fragment.

3.2 GENERATIVE MODEL APPROACH
We develop a Bayesian learning framework to tackle the wrapper adaptation and new attribute discovery problem. The probability for generating a particular text fragment \( (C, F, Y) \) can be expressed as follows

\[
P(C, F, Y | \alpha, \beta_p) = P(C|Y)P(F|Y; \beta_p)P(Y|\alpha)
\]

A particular set of \( \alpha \) and \( \beta_p \) refers the attribute generation control variable, Where \( \beta_p \) denotes the formatting feature generation variable of the page p. The expected log-likelihood function for the total number of text fragments N

\[
L(\alpha, \beta) = \sum_{i=1}^{N} \log P(C_i|Y_i = \alpha)P(F_i|Y_i = \alpha)\beta_p(i)
\]

We design an EM algorithm that approximates the parameters in (2) without knowing the actual value of the attributes of
Recall that in wrapper adaptation, we aim at extracting text fragments belonging to the attributes in $A_{\text{known}}$ and treat the attributes in $A_{\text{unknown}}$ and $\tilde{A}$ as a single component denoted as $\hat{A}$. Precisely, we estimate $\alpha^{t+1}$ as follows:

$$\alpha^{t+1} = \frac{\sum_{i=1}^{N} P(y_i = a|C_i, F_i; \alpha^t, \beta^{t+1}_p)}{N}$$  

On the other hand, $\beta^{t+1}_p$ controls the formatting feature using $P(F_i|Y_i, \beta^{t+1}_p)$.

$$\beta^{t+1}_p = \alpha^t \beta^t + \frac{1}{\sum_{i=1}^{N} \sum_{k=0}^{1} \sum_{a \in A} P(F_i = f_{(i)}|Y_i = a; \alpha^t, \beta^t, \chi(p, i))} = \{a; \beta^t\} P(Y_i = a|C_i, F_i; \alpha^t, \beta^t, \chi(p, i))$$

Where $k$ refers the formatting feature of the $i$th iteration, $\chi(p, i) = 1$ if $p(i) = p$ and 0, otherwise. For each attribute, the top $N$ useful text fragments with their probabilities belonging to this attribute higher than a certain threshold $\theta$ will be selected as the training examples for learning the new wrapper for the new unseen Web site.

### 3.3 NEW ATTRIBUTE DISCOVERY

The goal is to discover previously unseen attributes together with their semantic labels in the unseen sites. By Bayes theorem:

$$P(H|C, F, S; \alpha, \beta, \gamma) = \frac{P(Y = a|C, F; \alpha, \beta)P(S|H)}{P(C, F)}$$

Since $H$ and $Y$ are both unobservable, we drive the following expected log likelihood function

$$\sum_{i=1}^{N} \sum_{h=0}^{1} \sum_{a \in A} P(Y_i = a|C_i, F_i; \alpha, \beta, \gamma) P(S_i|H_i = h)$$

Next, the new attributes and the associated semantic labels can be discovered based on the estimated parameters. The E-Step and M-Step at the $t$th iteration are described as follows:

**E-Step**

$$P(H_i|C_i, F_i, S_i; \alpha, \beta, \gamma) \propto \sum_{a \in A} P(Y_i = a|C_i, F_i; \alpha, \beta) P(S_i|H_i)$$

**M-Step**

$$\nu^{t+1} = \arg \max_{\gamma} \log P(H_i|C_i, F_i, S_i; \alpha, \beta, \gamma)$$

Let $H$ be represented as a binomial distribution, $\nu^{t+1}_p$ for page $p$ is
\[
\gamma_i^{t+1} = \frac{\sum_{n=1}^{N} \sum_{e \in E} \sum_{k=1}^{S_i} P(Y_i \mid C_i, F_i; \alpha, \beta) P(S_{i,k} \mid H_i) P(H_i \mid Y_i = \alpha; \gamma_i^{t}(p, i)) \chi(p, i)}{R}
\]

Where \( S_{i,k} \) refers to the kth semantic label candidate for the ith useful text fragment; \( |S_i| \) refers to the number of semantic label candidates for the ith useful text fragments; and R refers to the total number of pairs of useful text fragments and semantic label candidates.

Lexical semantics [8], [miller 1999] and similarity measures are the meaningful relatedness that is not as intuitive and clear a notion as it may seem at first. Few lexical semantics are homonym, homograph, homophone, Polyseme and Capitonym. Let x and y represents two texts respectively that converted from Web documents, whose meanings are represented by the sets A and B. The lexical semantic Association (LSA) between x and y is given by

\[
LSA = m(A \cap B)/m(A \cup B)
\]

In which m indicates the number of words in a set.

According to fitness value, Chromosomes will be selected for further process of recombination.

4 EVALUATION ON WRAPPER ADAPTION

We conducted a set of experiments in each domain to evaluate our framework for wrapper adaptation. The extraction performance is evaluated by precision, recall, f-measure, true positive and false positive. Precision is defined as the number of items for which the system correctly identified divided by the total number of items it extracts. Recall is known as the number of items for which the system correctly identified divided by the total number of actual items. F1 measure, which is defined as the harmonic mean of recall and precision, is also used in the evaluation. True positive is defined as the intersection of items between the target items and observation. Finally false positive is the items belong to the observation and it does not present on the true positive.

5 EMPIRICAL RESULTS

We conducted extensive experiments on more than 10 real world Web sites collected from the book domain. The Web sites labeled with S1-S10 were collected from the book domain. For example, when adapting the wrapper from the source Web site S1 to the target Web site S2, the source wrapper is trained from the training examples manually annotated on the training set of S1. Wrapper adaptation is conducted to automatically identify training examples in the training set of S2. C1 and C2 are the clustered web pages based upon their structure and content. In each domain, we used different sets of parameters for adapting the wrapper from one of Web sites to another and vice versa.

Table 1 shows the performance of wrapper generation in book domain. Compared with ROADRUNNER, KNOWITALL, TEXTRUNNER the extraction performance of our approach is very promising. The performance of Bayes with EM approach is nearly the same as that of our approach in extracting the attribute in book domain. Table 2 shows that our new attribute discovery approach can be applied to each new unseen site to discover new attributes. For example, the wrapper learned from S1 is first adapted to S2-S10. Next, our new attributes are discovered contained in S2-S10.

<table>
<thead>
<tr>
<th>RoadRunner</th>
<th>KnowITAll</th>
<th>TextRunner</th>
</tr>
</thead>
<tbody>
<tr>
<td>P (%)</td>
<td>R (%)</td>
<td>F (%)</td>
</tr>
<tr>
<td>S1</td>
<td>51.2</td>
<td>62.2</td>
</tr>
<tr>
<td>S2</td>
<td>52.44</td>
<td>62.14</td>
</tr>
<tr>
<td>S3</td>
<td>52.84</td>
<td>63.25</td>
</tr>
<tr>
<td>S4</td>
<td>53.63</td>
<td>64.42</td>
</tr>
<tr>
<td>S5</td>
<td>54.02</td>
<td>63.35</td>
</tr>
<tr>
<td>S6</td>
<td>54.48</td>
<td>66.42</td>
</tr>
<tr>
<td>S7</td>
<td>54.64</td>
<td>66.94</td>
</tr>
<tr>
<td>S8</td>
<td>56.32</td>
<td>67.56</td>
</tr>
<tr>
<td>S9</td>
<td>58.72</td>
<td>68.44</td>
</tr>
<tr>
<td>S10</td>
<td>59.4</td>
<td>68.52</td>
</tr>
<tr>
<td>Avg</td>
<td>54.76</td>
<td>65.52</td>
</tr>
</tbody>
</table>

Table 1. Average performance of wrapper generation in book domain.
Table 1. (Continuation)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>P (%)</th>
<th>R (%)</th>
<th>F (%)</th>
<th>Tp (%)</th>
<th>Fp (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>72.2</td>
<td>61.4</td>
<td>66.36</td>
<td>71.23</td>
<td>28.72</td>
</tr>
<tr>
<td>C2</td>
<td>73.7</td>
<td>62.81</td>
<td>67.82</td>
<td>72.63</td>
<td>27.36</td>
</tr>
</tbody>
</table>

Table 2. New attribute discovery via new approach

<table>
<thead>
<tr>
<th>New attributes identified from unseen web sites</th>
<th>New attributes not identified correctly</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1 Book seller rating, New form, used form</td>
<td>Bn.com price, buy new</td>
</tr>
<tr>
<td>S2 Format, shipping</td>
<td>------</td>
</tr>
<tr>
<td>S3 Format, availability</td>
<td>Market place</td>
</tr>
<tr>
<td>S4 Review, shipping</td>
<td>------</td>
</tr>
</tbody>
</table>

6 CONCLUSIONS
In this paper, we proposed an adaptable wrapper generation to generate wrapper for domain specific web pages. Our approach can automatically adapt the information extraction patterns previously learned in a source Web site to new unseen sites, at the same time, discover new attributes together with semantic labels and lexical semantic search. Our method is fully automatic and it generates a reliable and accurate wrapper for web data integration purpose. Extensive experiments from more than 10 real world Web sites in three different domains were conducted to evaluate our method and the results prove the approach to be promising.

7 FUTURE WORK
For information extraction the immediate future looks bright. With the continued availability of cheaper processing and storage, integration of increasingly thorough IE techniques with other text-processing applications will continue. As a result, even for general applications, enterprises will have access to precise text search and retrieval capabilities without needing to integrate additional tools into their systems. More lexical semantics for the new attribute discovery can be applied. Furthermore the abbreviations and annotations can be considered for wrapper generation.

REFERENCES:


