Research Paper

Volume : 3 | Issue : 4 | April 2014 • ISSN No 2277 - 8179

Emulate Rule Extraction From Identical Web Sites Based on Rule Ontology

Engineering

KEYWORDS : Rule extraction, rule onto, semantic similarity, stemming.

Mrs. S. C. Punitha	Research Scholar,Department of Computer Science and Engineering,Karunya University,Coimbatore, India
Dr. P. Ranjit Jeba Thangaiah	Head ,Department of Computer Applications,Karunya University, Coimbatore, India.
M.Punithavalli	Dean, Sri Ramakrishna College of Arts and Science for Women.

ABSTRACT

Web documents are usually semi-structured, and manually processing it is difficult.To overcome the above problem the information contained existing web pages are converted into ontology.We proposed an automatic rule acquisition procedure using ontology, named Rule To Onto, that includes information about the rule components and their structures. We started from the idea that it will be helpful to acquire rules from a site if we have similar rules acquired from other similar sites of the same domain. Rule To Onto is a generalized, condensed, and specifically rearranged version of the existing rules. The rule acquisition procedure consists of the rule component identification step and the rule composition step. We used stemming and semantic similarity in the former step and developed the A algorithm[1] in the latter step. This paper focuses on using Rules Extraction to automatically augment web pages with semantic annotations, and find out whether and how ontology Rules Extractions could be used in the extractionprocess.*

I. INTRODUCTION

Automatic Rules possession is a relatively fresh area of work and in the literature there is still very little guidance about how Rules Extraction projects should be managed. Every project typically starts by setting its scope and goals; in case of Rules Extraction one needs to define the items to be extracted, from which documents to perform extraction, how to integrate and store the extracted data, and how to use it in a final application. Rule acquisition is as essential as ontology acquisition, even though rule acquisition is still a bottleneck in the deployment of rule-based systems. This is time consuming and laborious, because it requires knowledge experts as well as domain experts, and there are communication problems between them. However, sometimes rules have already been implied in Web pages, and it is possible to acquire them from Web pages in the same manner as ontology learning [2]. There are some problems with extracting rules from text. First, which words of the Web page are rule components and which types of rule components are they?. Second, how can we compose rules with the rule components? There are numerous possible combinations of making rules. Our idea for solving these problems is using rules of similar sites in limited situations under a couple of assumptions. Let us suppose that we have to acquire rules from several sites of the same domain. The sites have similar Web pages explaining similar rules from each other. A comparison shopping portal can be an example.

II. STEPS IN EXTRACTION OF RULES

2.1 Role of Rule Ontology

The purpose of using ontology in our approach is to automate the rule acquisition procedure. The starting point of our approach is that it will be helpful for acquiring rules from a site, if we have similar rules acquired from other similar sites of the same domain[3] . Rule ontology, which , includes the information about rules including terms, rule component types, and rule structures. We named the rule ontology Rule To Onto. It has the advantage that it is structured information and is much smaller than rule bases, so that it is easy to reuse, share, and accumulate. However, some part of the burden of rule acquisition is shifted to ontology acquisition in our approach. The fact that we need similar sites and rule bases is a significant constraint in our approach, even though we have an automatic procedure for building ontology from the existing rule bases.

The following figure shows the conversion of html document to text .

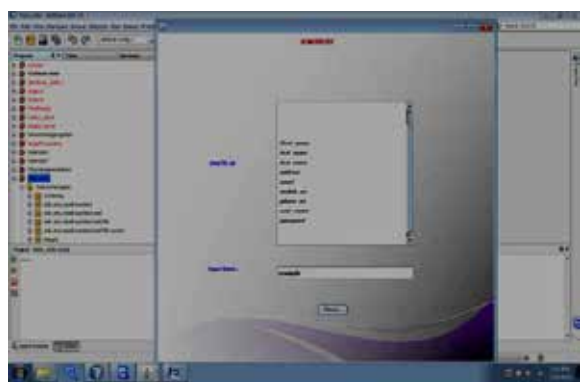

FIG.1 converting HTML TO TEXT

2.2 Rule Extraction

A large number of Rules Extraction systems are based on manually defined grammars whose aim is to identify segments of interest in the stream of processed text. These systems Rules Extraction the processed text using the phrase representation, Regular grammars have been the most popular since text can be searched very effectively for their matches using finite state automata. The following figure shows the extractions.

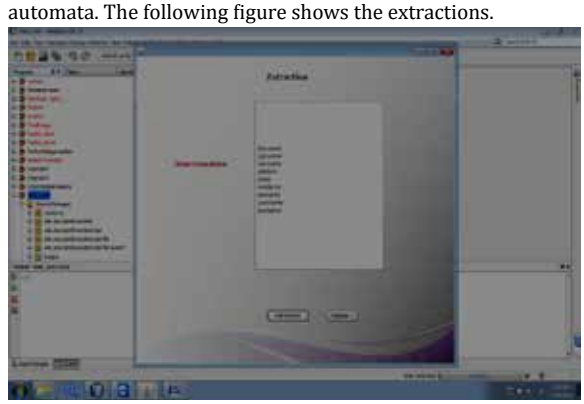

FIG.2 Extractions Using Rule Onto

Rules Extraction system utilized a cascade of regular grammars to capture occurrences of increasingly complex events in text; the latter layers matched output of the former. For example, the first levels of regular grammars captured multi-word expres-

sions, then noun and verb groups RULE ONTO Extractions and Rules Extraction, modeling parts of reality with domain RULE ONTO Extractions became increasingly popular and a number of ontology authoring tools appeared. Rules Extraction techniques became the natural choice to populate these ontology Rules Extractions with instances from text automatically.

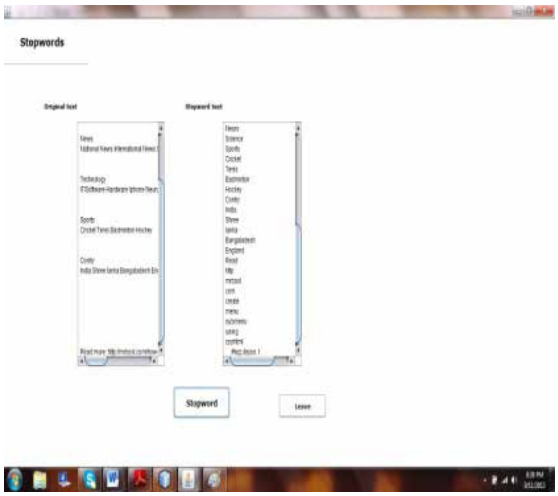

FIG 3Stop Word Removal Process

2.3 Stemming and Semantic Similarity

In information retrieval, stemming [4] is the process for reducing inflected (or sometimes derived) words to their stem, base or root form—generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. typically smaller list of "rules" is stored which provides a path for the algorithm, given an input word form, to find its root form. Some examples of the rules include:

- if the word ends in 'ed', remove the 'ed'
- if the word ends in 'ing', remove the 'ing'
- if the word ends in 'ly', remove the 'ly'

The following figure shows the stemming of the word "timing" to "time".

Fig 4 .Stemming

FIG 5 Filtering Processes

Even though the patterns and contents of rules of different sites are similar, they usually use different terms that have the same meaning. They use synonyms in most cases, but they sometimes use semantically similar[5] concepts with different rule structures. For example, Amazon uses the concept region for shipping destinations, but Powells.com uses country in every shipping rate rules. Country is not the synonym of region, but is semantically similar to region.

Therefore, we decided to use a semantic similarity measure in addition to synonyms in order to increase the recall rate when we identify variables and values

The following figure shows the semantic similarity of few keywords.

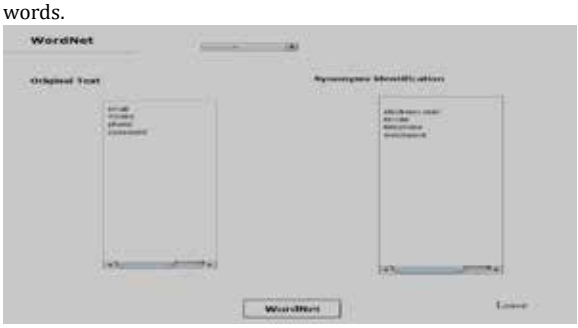

FIG 6.Semantic Similarity

III. RULE ONTOLOGY GENERATION

RuleToOnto is domain specific knowledge that provides information about rule components and structures. It is possible to directly use the rules of the previous system instead of the proposed ontology. However, it requires a large space and additional processes to utilize information on rules, while Rule To Onto is a generalized compact set of information for rule acquisition. Thus, we use Rule To Onto instead of the rules themselves. While the rule component identification step needs variables, values, and the relationship between them, the rule composition step requires generalized rule structures. Therefore, Rule To Onto represents the IF and THEN parts of each rule by connecting rules with variables with the IF and THEN relations, in addition to basic information about variables, values, and connections between variables and values. The Rule To Onto schema has three object properties Has Value, IF and THEN, and three classes, Variable, Value, and Rule.

3.1 RuleToOnto Generation Using Protégé

The following depicts the RuleToOnto Generation of the keywords using Protégé[6]: Email, mobile, phone, password

```
<?xml version="1.0"?>
<rdf:RDF
xmlns="http://a.com/ontology#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xml:base="http://a.com/ontology">
<owl:Ontology rdf:about=""/>
<owl:Class rdf:ID="email">
<rdfs:subClassOf>
<owl:Class rdf:about="#mobile"/>
</rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID="phone">
<rdfs:subClassOf>
<owl:Class rdf:about="#password"/>
</rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID="null"/>
<owl:Class rdf:ID="null">
<rdfs:subClassOf rdf:resource="#null"/>
</owl:Class>
<owl:Class rdf:ID="null">
<rdfs:subClassOf>
```

<owl:Class rdf:ID=”null”/>
</rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID=”null”>
<rdfs:subClassOf rdf:resource=”#null”/>
</owl:Class>
<owl:Class rdf:ID=”null”>
<rdfs:subClassOf rdf:resource=”#null”/>
</owl:Class>
<owl:Class rdf:ID=”null”>
<rdfs:subClassOf>
<owl:Class rdf:about=”#null”/>
</rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID=”null”>
<rdfs:subClassOf rdf:resource=”#null”/>
</owl:Class>
<owl:Class rdf:ID=”null”>
<rdfs:subClassOf>
<owl:Class rdf:about=”#null”/>
</rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID=”null”>
<rdfs:subClassOf>
<owl:Class rdf:about=”#null”/>
</rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID=”null”>
<rdfs:subClassOf>
<owl:Class rdf:about=”#null”/>
</rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID=”null”>
<rdfs:subClassOf rdf:resource=”#null”/>
</owl:Class>
<owl:Class rdf:ID=”null”>
<rdfs:subClassOf rdf:resource=”#null”/>
</owl:Class>
<owl:Class rdf:ID=”null”>
<rdfs:subClassOf rdf:resource=”#null”/>
</owl:Class>
<owl:Class rdf:ID=”null”>
<rdfs:subClassOf rdf:resource=”#null”/>
</owl:Class>
</rdf:RDF>
<!-- Created with Protege (with OWL Plugin 1.2 beta, Build 139) http://protege.stanford.edu -->

3.2 Precision and recall

Precision also called positive predictive value is the fraction of retrieved instances that are relevant, while recall also known as sensitivity is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

$$\text{Precision} = \frac{A}{A+C} * 100\ \%$$

$$\text{Recall} = \frac{A}{A+B} * 100\ \%$$

A:number of relevant terms retrieved
B:number of relevant terms not retrieved
C:number of irrelvant terms retrieved

The following figure shows the precision and recall chart of the extracted rules

FIG.5 Precision And Recall Chart

CONCLUSION

Methodology that utilizes the sets of extraction rules to extract the relationships and entities (information constructs) from unstructured text that validates and semantically associates the extracted constructs with the given domain ontology, and perform ontology enrichment with newly found relationships, whenever possible.

Methods to represent these construct using existing (non-extended) RDF specification.

Our subsequent work will be focused on not yet supported XML Schema components, so that more detailed and precise ontologies can be generated. Furthermore, to improve the support for document oriented XML (also with mixed content) by letting the user control the transformation process to get more influence on the mapping and also to reach an improvement of the performance when processing the OWL instances.

REFERENCE

[1] J.C. Beck and M. Fox, “A Generic Framework for Constraint Directed Search and Scheduling,” AI Magazine, vol. 19, no. 4, pp. 101-130, 1998. | [2] P. Cimiano and J. Volker, “Text2onto-a Framework for Ontology Learning and Data-Driven Change Discovery,” Proc. 10th Int’l Conf. Applications of Natural Language to Rules Systems (NLDB), pp. 227-238, 2005. | [3] Sangun Park and Juyoung Kang “Using Rule Ontology in Repeated RuleAcquisition from Similar Web Sites” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 6, JUNE 2012 | [4] B. Priyadharshini, A. Mohana, R. Radhika, “Retrieving Rules Using Composetoonto Based On Stemming Algorithm From Similar Web Sites” , International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 3, March - 2013 ISSN: 2278-0181 | [5] David Sánchez, Montserrat Batet, David Isern, Aida Valls, “Ontology-based semantic similarity: a new feature-based approach” Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA),Departament d’Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Avda. Països Catalans, 26. 43007 Tarragona (Spain) | [6] P. Buitelaar, D. Olejnik, and M. Sintek, “A Prote´ge´ Plug-in for Ontology Extraction from Text Based on Linguistic Analysis,” Proc. First European Semantic Web Symp. (ESWS), 2004. |