Automatic Rule Extraction from Frequent Rules In Similar Network Sites

Aakanksha Singh [#1], B.Muruganantham [*2]

1Dept of Computer Science and Engineering, 2 Asst.Prof [Sr.G] Dept. of Computer Science and Engineering

[1]SRM University, Chennai, India.

[2]SRM University, Chennai, India.

***Abstract*—Knowledge is an essential part of most Semantic Web applications and ontology. Ontology is the most important part of the knowledge. Ontology learning, which refers to extracting conceptual knowledge from several sources and building ontology from scratch, enriching, or adapting an existing ontology is one of the attempts at knowledge acquisition. Ontology is not sufficient to represent inferential knowledge. The ontology-based analysis with description logic is a popular issue of the Semantic Web. Rules are obtained from several sites of the same domain. The existing system has some problems in rule extracting. First, the web pages are identified for rule components and for their types. Second thing is how to compose the rules with rule components. The domain has similar Web sites explaining similar rules from each other. It decreases the burden on the knowledge experts and domain experts. Our idea for solving these problems is using rules of similar sites in limited situations. The two main steps of rule acquisition, which consists of rule component identification such as variables and values in Web pages by using RuleToOnto and rule composition with the identified rule components. We performed experiments demonstrating that our ontology-based rule acquisition approach works in a real-world application.**

***Index Terms*—Rule Acquisition, Breath first Search, Rule ontology.**

I. INTRODUCTION

The web is continuously evolving towards new improvements after going through Web 2.0 since it was created with hyperlinks and multimedia capabilities. The Semantic Web, which is the key component of Web, is an evolving development of the World Wide Web in which the semantics of information and services on the Web are being defined. This is enabling the Web to understand and satisfy the requests of people and machines to use the Web content. Knowledge is an essential part of most Semantic Web applications and ontology, which is a formal explicit description of concepts or classes in a domain of discourse, is the most important part of the knowledge. However, ontology is not sufficient to represent inferential knowledge. This is because ontology-based reasoning has limitations compared with rule based reasoning, even though ontology-based reasoning with description logic is a popular issue of the Semantic Web. That is, inferential rules are also the essential part of knowledge of the Semantic Web. Many attempts have been made at knowledge acquisition in order to obtain enough knowledge for Semantic Web applications. Ontology learning [1], which refers to extracting conceptual knowledge from several sources and building ontology from scratch, enriching, or adapting an existing ontology, is one of the attempts at knowledge acquisition. Most ontology learning approaches acquire knowledge from the Web, because it offers a large amount of valuable information for every possible domain. 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. That is, most of the rule components already exist in Web pages. It means that we can acquire rules more easily by using an automatic rule acquisition method rather than the old method with domain experts and knowledge experts. However, there are some problems with extracting rules from text for example, item is a variable and book is a value in. 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. The comparison of simple data such as book prices does not need rules, but delivery cost calculation with various options and applying free shipping rules and return policies needs rule. Therefore, the portal should acquire rules about delivery options, shipping rules, and return policies from shopping malls if it wants to provide an intelligent service comparing more than just prices. In this case, the portal should repeatedly acquire similar rules from multiple malls and the rules are very similar to each other in terms of their shape and content. For example, Amazon.com (in short Amazon) and

Barns&Noble.com (in short BN) posted similar Web pages explaining shipping rules and return policies, and the terms and rules implied in the Web pages are also similar. Let us suppose that we are going to acquire rules from BN after we have acquired rules from Amazon. We might be able to use the terms and structures of rules from Amazon in order to acquire rules from BN. By using the information about rules, we can develop an automatic rule acquisition procedure which decreases the burden on the knowledge engineer. However, there are assumptions in our approach. First, the rules across different sites should be similar. Our example fits this assumption. The second assumption is that the Web page should include practical and executable rules. The last assumption is that the target application requires rule-based reasoning. As we repeat the rule acquisition process across several sites, we can accumulate rules. However, as the size of rule base increases, it becomes hard to reuse rules. Therefore, we used an ontology named RuleToOnto, which represents the information about rules including terms, rule component types, and rule structures. The ontology cads crease the amount of information and reduce the work of utilizing the information in rule acquisition, because it is generalized and specifically rearranged for rule acquisition. Moreover, the ontology can be accumulated and reused throughout repeated rule acquisition. The main objective of our research is to propose a rule acquisition procedure that automates repeated rule acquisition [1] from similar sites by using the rule ontology RuleToOnto. We proposed two main steps of rule acquisition, which consists of rule component identification and rule composition with the identified rule components. In other words, we identify rule components such as variables and values in Web pages by using RuleToOnto [1] in the first step, and we combine the variables to compose rules in the second step. The rule component identification step and its evaluation were published in the previous work, and the concept of rule composition with an incomplete algorithm was introduced in a conference. After the conference, we completed and enhanced the rule composition algorithm, and conducted experiments to evaluate the performance. Therefore, the main contribution of this paper is the complete and detailed rule composition process with examples and its evaluation.

II. METHODOLOGY

The webpage holders first register with server. After, server sends the request to webpage holder such a user mentioned webpage. Then the web page holder sends the requested webpage to server. Afterwards server, shows the automatically extract the similar rules from the web page. Variables and values are collected from each HTML web page and identified relationship between them using RuleToOnto. After, shows the rule, identified variables and values. Now to assign instance to identified rules, variables and values in the identified rule paragraph. The Rule ordering, variable ordering and total ordering of rules are measured. That shows how to search for identify the similar variable instance and rule instance in the above measured elements. This is done by best first search method. First Variable Instances is to make an initial set. Because, extract variable instances that match the first variable of Total Order of rule candidate and put them into initial set. Continuously choosing up to the most suitable variable instance for the current variable from initial set with the evaluation function. Otherwise if the current variable is the last in Total Order of rule candidate gets the next variable of current Variable Instance from Total Order of rule candidate. It is empty when we assign all variable instances to the variables in Total Order of Rule Candidate If current Variable Instance is not the last variable, transfer it from Initial set to last set and update total order of rule candidate with recalculate variable count, variable ordering and rule ordering.

Subsequently, get the next variable from total Order of rule candidate and get of the variable. Every element of next variable instances should be an instance of the next variable of current variable Instance in total order of rule candidate and should not appear in the current path., because an instance has already been assigned to a rule if it appears in the path. The path of current variable instance returns the path from the root to current Variable Instance in the search tree. RuleToOnto general structure has three classes, Variable, Value, and Rule and three object properties has value, IF and THEN.

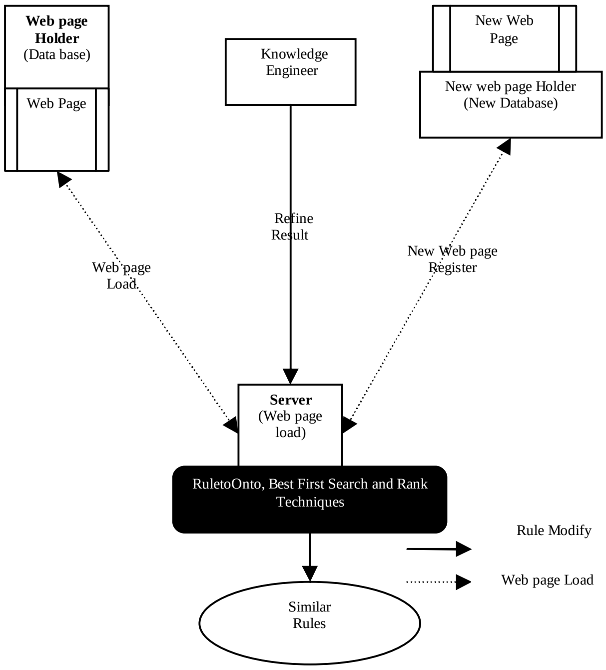
If the variable belongs to an IF part in the rule instance of RuleToOnto, assign the pair to the IF part of the rule. Otherwise, if it belongs to a THEN part, assign it to the THEN part. Show all similar rules in the page. Load another web site page and do the same process automatically. This module shows knowledge engineer refine the similar rules. After, these rules are ranked by server. Subsequently, forward the similar rules to newly registered website.

The existing Rule acquisition is still a bottleneck in the deployment of rule-based systems. The existing system is using knowledge experts as well as domain experts. Consume more time. Difficult to extract the rules from the web pages.

It is also need knowledge experts and domain experts. The contact problems between knowledge experts and domain experts. Our purpose is to extract inference rules from the text of Web pages. The proposed approach divided the rule acquisition procedure into two main steps. First step is rule component identification, proposed system identifies variables and values by using a Rule to onto that describes frequently used variables and values. Second step, the proposed system is composing rules from the identified rule components by using the rule structures of the ontology.

The advantages of all this approach are that the time consumption is reduced.Easy to extract the rules from the web pages. It does not need knowledge experts and domain experts.

III. ARCHITECTURE



The functional-Style syntax for rules can be made in this pattern.

```
EquivalentClasses(
   :Variable
    ObjectAllValuesFrom(:hasValue :Value)
)

EquivalentClasses(
   :Rule
   ObjectIntersectionOf(
      ObjectAllValuesFrom(:if :Variable)
          ObjectAllValuesFrom(:then :Variable)
    )
)
```

The identification variable instances can be denoted as some sort of patterns been acquainted.

R1 = {DS, RS, IT, RF}, R2 = {PF, IT, RT}

R3 = {RS, PF, IT, RT}, RC = {R1, R2, R3}.

Rule composition generates rule instances by assigning variable instances to rule candidates.

IV. RELATED WORK

Ontology [1], the core of semantic web, is usually limited by specific domain or created by given meaning passively. But, the generation of ontology is not only very difficult but also taking very long time. Therefore, the automatic generation of ontology takes very important part in semantic web. In this paper, we suggest the ways that make and renew the ontology, which are related with the keywords that users input in the search engine, automatically for the automatic generation of ontology that is not limited by specific domain. Input keyword and relation keyword becomes OWL, and the relation among the created OWL is expressed by Graph. The result is shown after the test of validity that W3C offers to verify the grammatical propriety of created OWL. The way of automatic generation of ontology that is being suggested, discharges generation and renewal of ontology automatically whenever searching is completed in the search engine. By the web based dictionary and thesaurus, the generation of ontology, for the new and related words, has become possible [2]. Business Intelligence (BI) requires the acquisition and aggregation of key pieces of knowledge from multiple sources in order to provide valuable information to customers or feed statistical BI models and tools. The massive amount of information available to business analysts makes information extraction and other natural language processing tools key enablers for the acquisition and use of that semantic information. We describe the application of ontology-based extraction and merging in the context of a practical e-business application for the EU MUSING Project where the goal is to gather international company intelligence and country/region information. The results of our experiments so far are very promising and we are now in the process of building a complete end-to-end solution [3]. In this paper we present Text2Onto, a framework for ontology learning from textual resources. Three main features distinguish Text2Onto from our earlier framework TextToOnto as well as other state-of-the-art ontology learning frameworks. First, by representing the learned knowledge at a meta-level in the form of instantiated modeling primitives within a so called Probabilistic Ontology Model (POM), we remain independent of a concrete target language while being able to translate the instantiated primitives into any (reasonably expressive) knowledge representation formalism. Second, user interaction is a core aspect of Text2Onto and the fact that the system calculates a confidence for each learned object allows designing sophisticated visualizations of the POM. Third, by incorporating strategies for data-driven change discovery, we avoid processing the whole corpus from scratch each time it changes, only selectively updating the POM according to the corpus changes instead. Besides increasing efficiency in this way, it also allows a user to trace the evolution of the ontology with respect to the changes in the underlying corpus [4]. The

World Wide Web is a vast source of information accessible to computers, but understandable only to humans. The goal of the research described here is to automatically create a computer understandable knowledge base whose content mirrors that of the World Wide Web. Such a knowledge base would enable much more effective retrieval of Web information, and promote new uses of the Web to support knowledge-based inference and problem solving. Our approach is to develop a trainable information extraction system that takes two inputs. The first is an ontology that defines the classes (e.g., company, person, employee, product) and relations (e.g., employed by, produced by) of interest when creating the knowledge base. The second is a set of training data consisting of labeled regions of hypertext that represent instances of these classes and relations. Given these inputs, the system learns to extract information from other pages and hyperlinks on the Web. This article describes our general approach, several machines learning algorithms for this task, and promising initial results with a prototype system that has created a knowledge base describing university people, courses, and research projects [5].

V. CONCLUSION

Inferential rules are as indispensible to the Semantic applications as ontology. An automatic rule extraction procedure using ontology, named RuleToOnto, which includes the related information about the rule components and their structures. It will be more convenient to acquire rules from a site if we have similar rules acquired from other similar sites of same domain. This work can be further enhanced as first to extend the rules extraction into various domains and second the rules can be classified as positive and individual rules. These positive rules are forwarded when new websites are added in a particular related domain.

REFERENCES

[1] Sangun Park and Juyoung Kang, “Using Rule Ontology in Repeated Rule Acquisition from Similar Web Sites”, vol. 24, No.6, June2012.

[2] Joon Shim and Hongchul Lee, “Automatic Ontology Generation Using Extended Search Keywords” Department of Information Management Engineering, Korea University, Seoul, Korea.

[3] Horacio Saggion and Adam Funk, “Ontology-based Information Extraction for Business Intelligence,” University of Shefield Regent Court, 211 Portobello Street, United Kingdom.

[4] Philipp Cimiano, Johanna VÄolker, “Text2Onto A Framework for Ontology Learning and Data-driven Change Discovery,” Institute AIFB, University of Karlsruhe.

[5] Mark Craven a Dan DiPasquo a Dayne Freitag, “Learning to Construct Knowledge Bases from the World Wide Web,” Carnegie Mellon University, 5000 Pittsburgh, PA, 15213-3891, USA bJust Research4616 Henry Street, Pittsburgh, PA, 15213, USA.