Two-Step Approach for Acquiring Semantic Relations from Textual Web Content

Sonal P.Patil
Department of Computer Science and Engineering
G.H.Raisoni Institute of Engineering and Management , Jalgaon

Rujata N.Saraf
Department of Computer Science and Engineering
G.H.Raisoni Institute of Engineering and Management , Jalgaon

Abstract - *Semantic*, is one of the most important & wide spread category of Natural Language Processing, related to study of meanings of a particular word in different context. With a vast growth of a World Wide Web, we face an increasing amount of information resources. Mining semantic relations from such a vast resources is quite difficult & different task from a normal text mining. The normal Text mining techniques are not sufficient for the knowledge discovery as these techniques simply transforms the free text into a group of words representation and hence does not preserve any semantics. In this paper, we are presenting two-step procedure to mine semantic relations from a textual web content. The procedures are – *RDF – Resource Description Framework* And *GP-Close – Generalized Pattern mining algorithm. RDF* is language specification, used to extract a metadata in the form of RDF statements representing semantic relations from raw data. For this purpose Natural Language processing techniques i.e. Myriad will be used. Once the metadata representing semantic relations is extracted, a novel *Generalized association Pattern mining algorithm (GP-close)* will be applied to discover the different association patterns from the RDF metadata. While finding such association patterns, the redundant over generalized patterns are eliminated by *generalization closure*, a term adopted by GP-Close algorithm. Finaly by eliminating redundant over generalized patterns, an accurate semantic relation will be extracted for knowledge discovery.

Keywords - text mining, Semantic Relation Extraction(RDF), Association Rule Mining

I. INTRODUCTION

The World Wide Web can be seen as one of the largest databases in the world. We usually have to face an increasing amount of information resources, as there is vast growth of the World Wide Web. The information Resources in WWW are mostly represented in free text. As text data are inherently unstructured and difficult for computer programs to directly procees, different techniques have been evolved such as Tex Mining [1]. With the help of such a techniques, users can easily get an expected data from the Web containing huge amount of unstructured data. Text-mining generally refers to the process of extracting interesting and non-trivial information and knowledge from unstructured text. An important difference with search is that search requires a user to know what they are looking for while text mining attempts to discover information in a pattern that is not known beforehand.

As Text Mining is all about analyzing unstructured information and extracting relevant patterns and characteristics, it has been consider as suitable method to give a structure look to unstructured information. Text mining technologies normally uses two subtasks [2] : *Text Refining and Knowledge Distillation*. *Text Refining* transforms free text into an intermediate representation form which is machine-process able, where as *knowledge distillation* discovers patterns or knowledge from the intermediate form. Existing techniques mainly transform text documents into simplistic intermediate forms, e.g., term vector and bags of keywords related to term vector. In this terms lose their semantic relations and texts lose their original meanings, as terms are treated as individual items in such simplistic representations. However, the relations that depicting the conceptual roles semantically, are lost in such a bag-of-keyword representations. Therefore, the original meanings of both the term & text cannot be differentiated against any more. And hence text mining techniques can only discover shallow patterns, such as term associations, deviations, and document clusters on the basis of such simplistic representation of text, which are statistical patterns of terms, not knowledge about text semantics. In this paper, we present two-step approach based on NLP techniques to overcome the limitation of text mining technologies to discover knowledge based on the detailed meanings of the

text. For this purpose, an intermediate representation that expresses the semantic relations between the concepts in texts is necessary, which can be done by the 2-step approach we presented in this paper..

II. PROPOSED WORK

For the purpose of knowledge discovery, in Semantic Web Mining the main focus is on extracting a semantic relations instead of extracting a bags of key-words related to the each and every word in user's search query. And to accomplish this task, the proposed two steps are as follows :

A. RDF (Resource Description Framework) :

Resources Description Framework (RDF), proposed by the World Wide Web Consortium (W3C), is a language specification to model an unstructured data on web in to a machine-processable and human- readable semantic metadata, which is then used to describe Web resources on the Semantic Web[3]. In general, it's an ontology that provides a mechanism to capture information about the objects and the relationships that hold between them in some domain of interest. And hence It is a graphical language used for representing information about resources on the web. Resources on web are described in terms of properties and property values using RDF statements [6]. *RDF statements* are the basic element of RDF, which are usually triplets in the form of :

< *Subject, Predicate, Object* >

An RDF statement can express that there is a relation (represented by the predicate) between the subject and the object. The resources on the web in Semantic Web Mining can also be represented using a Conceptual graphs, with which the RDF can be interworkable. Conceptual Graphs serve as an intermediate language for translating natural languages into computer-oriented formalisms. As the full conceptual graph standard is complex for large scale applications, simplified conceptual graphs are used in many existing practices.

For example : The simplified conceptual graph for the sentence " *France Defeated Italy in the World Cup Quarter Final* " will be as follows :

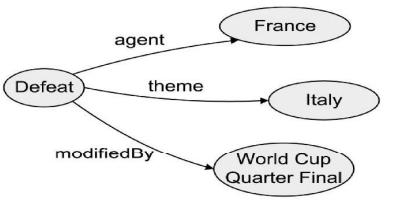

Figure 1 : Simplified Conceptual Graph

As shown in diagram, Each directed arc in the simplified conceptual graph as a semantic relation consisting of a subject (the start node of the arc), a predicate (the label or type of the arc), and an object (the end node of the arc). Each of these relations can be encoded using an RDF statement[4]. With the help of such a conceptual graph we can extract number of Semantic relations in the form of RDF statements.

<Defeat, agent, France>,
<Defeat, theme, Italy>, and
<Defeat, modifiedBy, World Cup Quarter Final>.

This is the simplified view of conceptual graph, The full conceptual graph is representated by following a predefined notations and various Events and states.

For example : consider the sentence - *If a farmer owns a donkey, then he beats it* , and the full conceptual Graph for this sentence is as shown in figure. A representation that treats events and states as entities linked to their participants by *case relations* or *thematic roles*.[5]

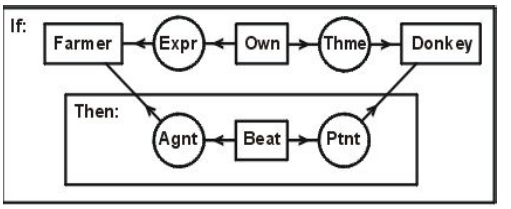

Figure 2 : Full Conceptual Graph

The figure shows the state of owning linked to its participants by the relations *experiencer (Expr)* and *theme (Thme)*, and the act of beating by the relations *agent (Agnt)* and *patient (Ptnt).* After extracting all possible semantic relations from the conceptual graph, all the extracted semantic relations will be stored in RDF metadata, on which a GP-Close Algorithm will be applied in second step to reduce the redundant overgeneralized patterns in relation pattern search space.

B. GP Close Algorithm :

GP-Close is nothing but *Generalised association Pattern Mining Algorithm,* is applied to discover the underlying relation association patterns on RDF metadata. For the discovery of such a underlying association relation patterns, *Association Rule Mining* have been used. At the end of this process it might be possible that there can be more than one relation having same meaning. Extraction of Such a sentences which are having same meaning, are called as Over generalized pattern . For pruning the large number of such redundant over generalized patterns in relation pattern search space, the GP-Close algorithm adopts the notion of generalization closure for systematic overgeneralization reduction. The GP-Close algorithm is always used with efficient *pattern space pruning* and full *overgeneralization reduction* techniques on any RDF metadata to reduce the redundant semantic relations which is helpful to mine exact relation from RDF metadata. Following are the main techniques use in GP-Close algorithm :

❖ *Association Rule Mining on Textual Web :*

Association Rule Mining has become one of the key data mining techniques in the field of Knowledge Discovery in Database (KDD). The Association Mining Rule is stated as follows : "Given a set of items I and a large database of transactions D, where each transaction is a set of items $T \subseteq I$ with a unique identifier t_{id} , an association rule is an implication of the form $X \Rightarrow Y$, where $X,Y \subseteq I$ (called itemsets or patterns) and $X \cap Y = \emptyset$. A transaction T supports an itemset X if $X \subseteq T$. The support of an itemset X, denoted by *supp(X),* is the fraction of transactions in D that support X. Additionally, the support of an association rule $X \Rightarrow Y$, denoted by *supp(*$X \Rightarrow Y$*)* , is defined as *supp (*$X \cup Y$*).* Then the confidence of $(X \Rightarrow Y)$, denoted by *conf(*$X \Rightarrow Y$*),* is defined as :

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X).}$$

The problem of association rule mining is to discover all rules that have supports and confidences greater than some predefined minimum support (*minsup*) and minimum confidence (*minconf*). Mining association rules consists of two subtasks. The first task, known as *frequent itemset mining* (or frequent pattern mining), generates all itemsets that have supports higher than a minimum support (*minsup*) threshold. In the second task, association rules are generated based on the discovered frequent patterns [7]. Text databases cannot be efficiently analyzed by standard association mining algorithms. This is because the characteristics of text databases are quite different from those of relational and transactional databases. First, the number of distinct words in a text database is usually quite large (large size of I) and so is the number of words in each document (long transactions). The large number of words implies a large number of possible patterns (sets of words) both in a document collection and in each individual document. Thus, a text AR mining algorithm needs to explore a much larger search space to find interesting patterns. Moreover, the document frequency of each word is usually very low [8]. So normally a Low *min sup* is used for mining text association rules. However, this will cause a large set of trivial patterns discovered. At the knowledge discovery stage, for discovering more detail knowledge, the conceptual graphs are first clustered into a hierarchy. Then,

pattern mining techniques, such as association rule mining, can be applied to this hierarchical structure. This method shows that meaningful and detailed patterns can be discovered from text using the conceptual graph representation.

C. Knowledge Discovery Process :

Using RDF as the intermediate representation, the *Knowledge Discovery Process* consists of two stages , that are : *Semantic Relation Extraction* and *Association Rule Mining.*

❖ *Semantic Relation Extraction :*

In the semantic Relation Extraction stage, various Natural Language Processing techniques are used to process unstructured text document from web. These NLP techniques which are used to process an unstructured textual data, are called as "Myriad techniques" . Text documents are processed using a myriad of natural language processing (NLP) techniques, such as pronominal coreference resolution, part-of-speech (POS) tagging, and sentence structure parsing. The semantic relations (composing the conceptual graphs) are then extracted from the tagged text sentences by using A set of predefined syntactic patterns . The extracted relations are encoded in RDF statements. In addition, a term taxonomy is constructed on the fly based on WordNet and domain-specific lexicons. The term taxonomy, described using RDF Schema [9] (a vocabulary specification for RDF), is, in turn, used in the subsequent stage.

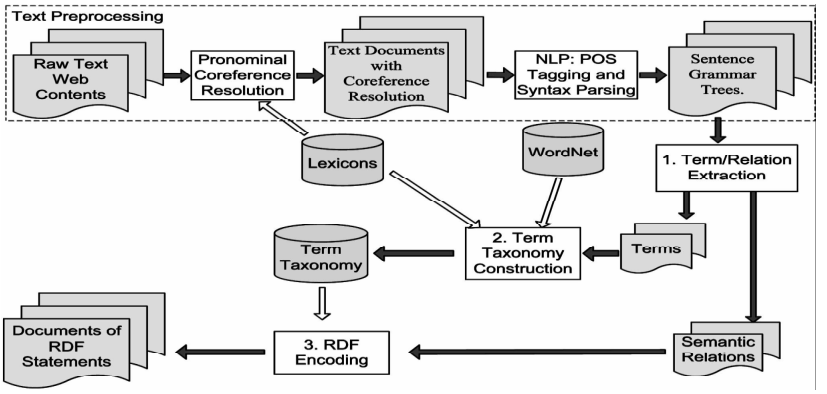

Figure 3 : Procedure for Sematic Relation Extraction

As shown in Figure 3, the raw textual Web content is sequentially preprocessed by a pronominal coreference resolution module including Lexical Analysis, POS (part-of-speech) tagging and syntax analysis (parsing) module. To eliminate the ambiguities of the pronouns in text, the pronominal coreference resolution function is used . In addition, domain-specific name lexicons are embedded for identifying name entities (NE) in text. After coreference resolution, each pronoun without having any ambiguity is replaced with the origin term that it refers to. The text documents are then tagged and parsed by two NLP tools, namely, part-of-speech (POS) tagger and parser[10] . After preprocessing, each parsed document contains a set of sentence grammar trees. Based on the sentence grammar trees, simplified conceptual graphs containing semantic relations are extracted. All these conceptual graphs are then encoded with the help of following modules.

1. Term and Relation Extraction :

Based on the preprocessing results, a set of predefined rules are used for extracting semantic relations from the sentence grammar trees. While extracting semantic relations, the important terms describing the major concepts are identified, such as, noun phrases (NP) and verb phrases (VP), in the grammar trees, followed by major three types of relations between these term [10]. The three relation types are :

- **< A; Agent; B > :** where A can be a VP and B can be an NP/VP. The relation indicates that B is the agent that performs the actionA.
- **< A; Theme; B > :** where A can be a VP and B can be an NP/VP. The relation indicates that B is the theme (i.e., recipient, object, or target) of the action A.
- **< A; Modified By; B > :** where A can be an NP/VP and B can be an NP/VP. This relation indicates that A is modified by B through a proposition.

2. Term Taxonomy Construction :

The terms (NP/VP) extracted from the sentence grammar trees are incrementally clustered into a term taxonomy with the assistance of WordNet . The atomic clusters in the term taxonomy are groups of synonyms. When a new term is inserted into the term taxonomy, first the search has been made to find whether there is a synonym group it can join. If there is such a synonym group, simply insert the new term into the synonym group and the structure of term taxonomy will not change; otherwise, it will be inserted as a new synonym group and the structure of the term taxonomy will change.

3. RDF Encoding :

In this module The term taxonomy and the semantic relations extracted from the sentence grammar trees are encoded as an RDF vocabulary and RDF statements, respectively.

❖ *Association Rule Mining* :

During the association rule mining stage, relation association patterns are discovered from RDF metadata which is extracted from the textual web. A problem for mining semantic relations is that a relation is might be repeated in many documents. Therefore, statistically significant patterns can hardly be extracted. To overcome this limitation, generalizations of the semantic relations are needed which is carried out by using various relation generalization algorithms. Such algorithms are only designed for mining patterns on atomic items, not on relations. Therefore, the existing methods cannot be directly applied to mining generalized relation patterns from RDF metadata. Even if we treat each relation as an item, the existing algorithms do not work efficiently on the RDF data. Hence Generalizing a semantic relation is complex as the relation can be generalized in many different ways. This implies that the generalized pattern search space can be very large and there can be many redundant overgeneralized patterns . Using the existing generalized pattern mining approaches, all generalized patterns including the overgeneralized ones will be extracted. It is not only computationally inefficient but also causes much redundancy in the mining results.

Therefore, the Generalized Association Pattern Mining Algorithm, called *GP-Close* (Closed Generalized Pattern Mining) is used to discover the generalized association patterns of semantic relations from RDF metadata. The large number of redundant overgeneralized patterns in relation pattern search space, are reduce by using *generalization closure* (notion adopted by GP-Close Algorithm) for systematic overgeneralization reduction.

III. CONCLUSION

In this paper we have present a two-step approach (i.e *RDF & GP-Closed Algo*) for mining semantic relations from textual web content. The web resources have huge amount of unstructured data and Mining such unstructured data with the help of normal text mining technique gives a flat bags of keywords & hence fails to preserve a Knowledge Discovery. The proposed seminar is all about *how the knowledge will be discovered* by mining semantic relations instate of mining Unstructured web data directly. For the purpose of Knowledge Discovery, various NLP techniques along with semantic Relation Extraction and Association Rule Mining techniques have been used.

REFERENCES

[1] [Berry Michael W., (2004), "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", *Springer Verlag, New York, LLC, 24-43.*

[2] Vishal Gupta, Gurpreet S. Lehal," A Survey of Text Mining Techniques and Applications", *Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, August 2009.*

[3] T. Berners-Lee, J. Hendler, and O. Lassila, "Semantic Web," *Scientific Am., vol. 284, no. 5, pp. 35-43, 2001.*

[4] Olivier Gerb_e, Guy W. Mineau, and Rudolf K. Keller,"Conceptual Graphs, Metamodeling and Notation of Concepts"

[5] Madalina Croitoru, Kees van Deemter, "A Conceptual Graph Approach to the Generation of Referring Expressions", *in IJCAI-07.*

[6] W3C OWL Web Site – http://www.w3.org/2004/OW/

[7] R. Agrawal, T. Imielinski, and A.N. Swami, "Mining Association Rules between Sets of Items In Large Databases," *Proc. ACM SIGMOD Conf., 1993.*

[8] Helen m. Meng,Member, IEEE, and Kai-chung siu, "Semantic Acquisition of Semantic Structure for Understanding Domain Specific Natural Language Queries", *IEEE Transaction of Knowledge and Data Engineering, Vol. 14, No.1, January/February 2002.*

[9] W3C, W3c RDF Schema Specification, http://www.w3.org/TR/rdf-schema/, 2005.

[10] T. Berners-Lee, J. Hendler, and O. Lassila, "Semantic Web," *Scientific Am., vol. 284, no. 5, 2001.*