Knowledge Extraction for Semantic Web

Gautam R. Raithatha

Information Technology
Shantilal Shah Engineering College, Bhavnagar, India
grraithatha@gmail.com

***Abstract*–Semantic web is the web with a meaning that computers can understand. In order to create a web with semantics, the information available from the unstructured or semi-structured web data has to be extracted and converted to a structured form that can be interpreted by computers. Different web mining techniques are used for extracting useful information from web data. In this paper, the main focus is to extract the concepts and conceptual relationships from unstructured textual data using web content mining to create the ontology. Ontologies can then be used to better serve the user queries.**

***Index Terms*–Web content mining, semantic web, ontology, knowledge (or information) extraction**

I. Introduction

There is a large amount of digital data present on the World Wide Web (WWW) in different forms such as web documents, files, media etc. Most of these data is not in structured form. Mostly these unstructured data is in textual format which only humans can understand while computers cannot interpret its meaning. Such textual data is not much descriptive. The important missing element in such documents is semantics [1]. Hence specialized knowledge services may require tools to be able to search and extract specific knowledge directly from the unstructured text on the Web [2], and convert it into a structured form that machines can interpret.

Web documents can be categorized into three categories based on their structure viz. un-structured, semi-structured and fully-structured. The un-structured document includes a web page or a file on web containing pure textual data without any formatting and with no other information. The semi-structured documents are the web pages which contain textual data with some structure such as HTML heading and paragraph tags which are useful for identifying hierarchical structure. The fully-structured documents contain information into structured form such as an XML or tabular form.

Web mining makes use of the data mining techniques to extract useful information from the web data. Web data includes web documents or web pages, hyperlinks between documents, website usage logs etc. Web mining can be classified into different types such as Web Content Mining, Web Structure Mining and Web Usage Mining. We will discuss each of it in Section II of this paper.

Semantic web, as the name implies, is the web with a meaning. Semantic web is considered as the second generation of World Wide Web. The backbone of semantic web is ontology [1]. Ontology is a formal representation of collection of concepts and their relationships [1]. Ontology is understandable to both machines and humans. Through ontology, meaning can be assigned to the web i.e. the semantic web can be constructed. We will discuss more about ontologies in Section IV of this paper.

Knowledge (or Information) extraction deals with identifying concepts or conceptual relationships from the unstructured data. Knowledge can be extracted from textual data, image, audio or any other type of data and stored in a structured form into the knowledge base of the search engine server. This extracted knowledge is used further to better serve the user queries.

Section II describes the basic concepts of web mining and its types. Section III describes the semantic web and its need. Section IV contains different definitions of ontology contained in different literatures and gives idea about web ontology language (OWL) and their need to add semantics to the web. Section V gives basic idea about knowledge extraction from unstructured textual data and illustrates the knowledge extraction procedure. Finally Section VI concludes the research paper.

II. Web Mining

Data mining is a field of computer science that deals with the knowledge discovery from the large databases by discovering patterns from the past data. Data mining extracts information from the databases and transforms it into understandable structure that can be used further for future prediction or any other use.

Web mining is the application of data mining techniques to extract knowledge from the web data, including web documents, hyperlinks between documents, websites usage log etc. [3]

The difference between data mining and web mining is that data mining operates on data from databases and finds useful patterns from it while web mining operates on web data (web documents, hyperlinks, logs etc.). In data mining the raw data is already in structured form which is used for future predication, while in web mining mostly the raw data is unstructured or semi-structured, which is converted to structured form for knowledge extraction.

There are three different types of web mining techniques viz. Web Content Mining, Web Structure Mining and Web Usage Mining.

A. Web Content Mining

As the name implies, in web content mining, the information is extracted from the contents present on web i.e. the web documents. Web documents may include different type of contents such as text, images, audio, video, tabular data etc. In web content mining, most research has been carried out in extracting knowledge form textual data. Web content mining also makes wide use of other technologies such as Information Retrieval (IR) and Natural Language Processing (NLP). Also recently there has been significant work done on extracting information from images which is a part of image processing field.

B. Web Structure Mining

In web structure mining, the main focus for mining is on structure of web data. If we consider the web data as a graph, then web pages becomes the nodes of the graph and the hyperlinks between these web pages becomes the edges of the graph which connects different nodes i.e. web pages. Web structure mining deals with discovering the structural information from the web. This structure can be further classified into two different types viz. Hyperlinks and Document Structure. Hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. Document structure specifies the structure of the content within a web page. The content can be organized in a tree-structured format, based on various HTML and XML tags within the page [3].

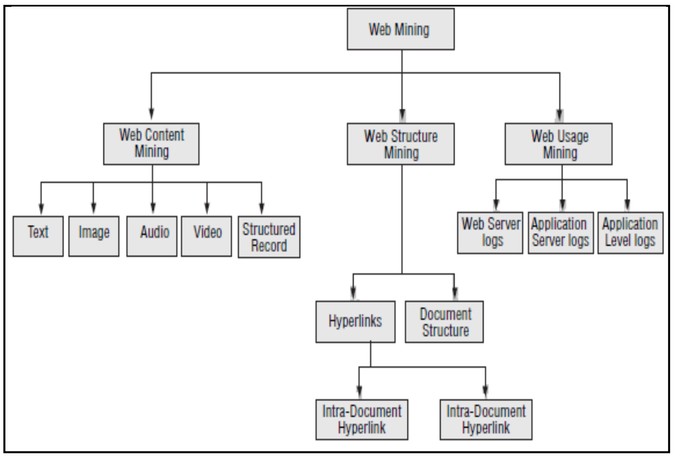


Fig. 1. Web mining taxonomy [3]

C. Web Usage Mining

Web usage mining, as the name implies, deals with discovering the web usage patterns of the users from the web usage logs of websites. Web usage mining helps to understand the browsing patterns of the users and can be used to better serve the needs of web-based applications and to make any necessary modifications for future. Usage data used for mining may include information such as user's identity, their location, browsing patterns etc. All these information can be obtained from website's usage log. The web usage data obtained from different sources can be classified into different types viz. Web Server Data, Application Server Data and Application Level Data. In web server data, user logs are collected by web server which includes information such as IP address, page reference, access time etc. [3] Application server data tracks various kinds of business events and logs them. Application level data include those events that are defined within the application and are logged by application itself which is programmed by the author.

Fig. 1 shows different types of web mining and their classification.

III. SEMANTIC WEB

Semantic Web is a collaborative movement led by the international standards body, the World Wide Web Consortium (W3C). Semantic web, as the name implies, is the web with a meaning. It is an extension of the existing World Wide Web. It provides a standardized way of expressing the relationships between web pages, to allow machines to understand the meaning of hyperlinked information [4]. The term "Semantic Web" was coined by Tim Berners-Lee, the inventor of the World Wide Web and director of the World Wide Web Consortium (W3C), for a web of data that can be processed by machines [5].

As discussed earlier in this paper, computers cannot understand the unstructured textual data. Semantic web proposes to help computers to read and interpret the data. The idea behind it is to add the metadata to the web pages. Although this will neither add any artificial intelligence to computers, nor it will make the computers self-aware, but it will give machine tools to find, exchange and interpret information to some extent [6]. Semantic web aims to convert the current web which includes unstructured or semi-structured documents, into a "web of data" by including the semantic content into web pages. Many of the websites are already using the semantic web, while there are a lot of the tools that are under development.

IV. ONTOLOGY AND WEB ONTOLOGY LANGUAGE (OWL)

Ontology can be considered as the backbone of semantic web. There are different definitions for ontology provided by different literatures. Some of the common definitions are:

(i) Ontology is a formal representation of collection of concepts and their relationships [1].
(ii) Ontology is an explicit specification of conceptualization [7].
(iii) Ontology is a term in philosophy and its meaning is "theory of existence" [8].
(iv) Ontology is a body of knowledge describing some domain, typically common sense knowledge domain [8].

Ontology is understandable to both machines and humans. Through ontology, meaning can be assigned to the web i.e. the semantic web can be constructed. Ontology creation is a semi-automatic process. The information extracted from the unstructured or semi-structured data forms an ontology which is in structured form, and is inserted into the knowledge base. The information present in the knowledge base is used for web mining, which improves the result of the user's query.

The Web Ontology Language (OWL) is a family of knowledge representation languages or ontology languages for authoring ontologies or knowledge bases. The languages are characterized by formal semantics and RDF/XML-based serializations for the Semantic Web [9]. In other words, the OWL Web Ontology Language is intended to provide a language that can be used to describe the classes and relations between them that are inherent in Web documents and applications [10].

V. Knowledge Extraction

Most of the information on the World Wide Web is in unstructured form. Our aim is to convert it into more structured form for creating semantic web i.e. to extract the knowledge from textual data and present it in a form that computers can understand. Thus when user queries for some information on the web, then only the relevant results are returned based on the knowledge extracted.

For example, suppose a user queried for some information such as the birthday of some well known person and that information is present as an unstructured text on some web page. The main focus here is that the required information, i.e. the birth date in this case, should have been stored in a structured form such as ontology or XML in the Knowledge Base (KB) of the search engine. Such structured form has to be created by extracting knowledge from the text so as to return the relevant information from the knowledge base when queried.

However, populating the knowledge base by extracting information from textual data to structured form is the main challenge. Ontology creation is a semi-automated process. It requires much of the time, effort and domain knowledge for manual work. There are many techniques and tools developed for information extraction from textual data. They are based on predefined templates and pattern-based extraction. However, the content on the web uses limitless vocabularies, structures and composition styles for defining approximately the same content, making it hard for any information extraction technique to cover all variations of writing patterns [2]. Also just extracting entities form textual content is not enough; it is also needed to identify the relationship between them.

A. Extraction Procedure

For knowledge extraction, the unstructured text from the web page is divided into paragraphs, which in turn are broken down into sentences. Each of these sentences is syntactically and semantically analyzed for information extraction. The Apple Pie Parser is used for grouping grammatically related phrases as the result of syntactical analysis. Semantic examination then locates the main components of a given sentence (i.e. ‘subject’, ‘verb’, ‘object’), and identifies the entities. This examination is done with the help of guiding tools such as WordNet, a general-purpose lexical database and GATE, an entity-recognizer. [2]

For example, consider the following sentence:

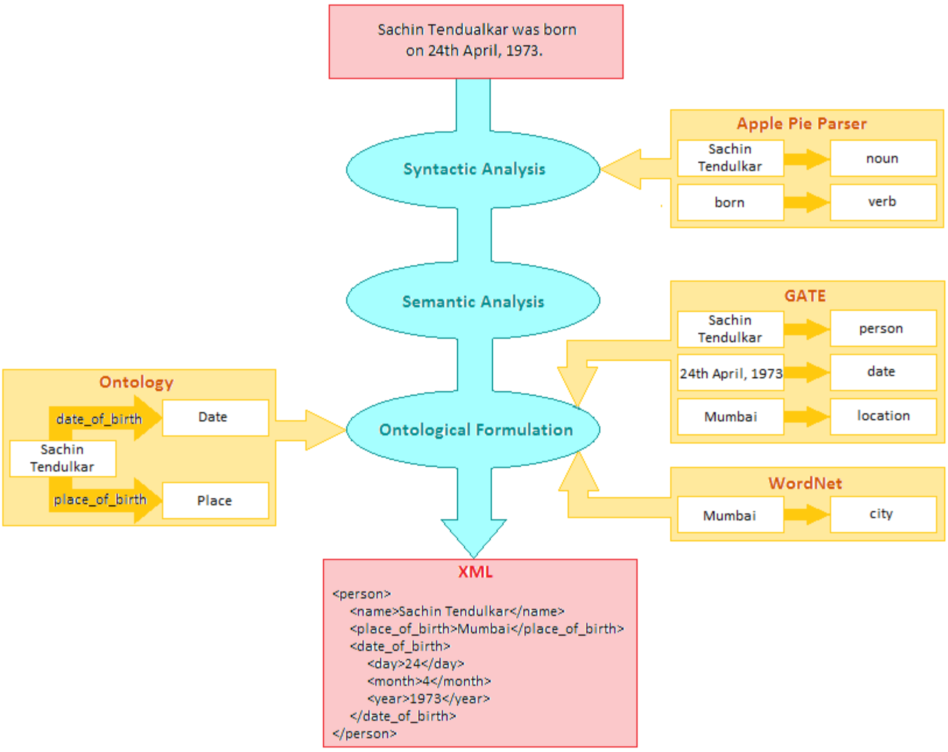
“Sachin Tendualkar was born on 24th April, 1973 in Mumbai.”

Humans can easily interpret the meaning of above sentence by reading it, but for the computer or any other machine, it is merely a text string consisting of letters and numbers, or at a lower level, binary stream (as computers store all data in binary form). To add semantics to the above sentence so that computers can also understand it, we need to construct ontology from it.

The challenge here is to extract the entities and binary relationship between them. If the three entities, “Sachin Tendulakar”, “24th April, 1973” and “Mumbai” are identified from above sentence as a person, a date and a location respectively, then too there are no relationships between them yet. They are just different entities extracted from a sentence. Knowledge about the domain is required to decide which relations are required and expected between the entities. What is needed is identifying a relationship which states that the second entity is a birth date of first entity i.e. the “date_of_birth” relationship is needed and third entity is the birth place of first entity i.e. “place_of_birth” relationship is needed.

Fig. 2 shows the extraction result for the above sentence. The GATE and WordNet tools are used to obtain annotations such as “Sachin Tendulkar” is a person’s name, “24th April, 1973” is a date and “Mumbai” is a location. The following two relationship triples are identified form above entities [2]:

Sachin Tendulkar – date_of_birth – 24th April, 1973
Sachin Tendulkar – place_of_birth – Mumbai



```
<person>
   <name>Sachin Tendulkar</name>
   <place_of_birth>Mumbai</place_of_birth>
   <date_of_birth>
      <day>24</day>
      <month>4</month>
      <year>1973</year>
   </date_of_birth>
</person>
```

Fig. 2. An example of knowledge extraction [2]

The output of the above information extraction process is the XML representation of the concepts and the conceptual relationships. This XML is then sent to the ontology server for inserting into knowledge base. When a user now queries for birth date or birth place of Sachin Tendulkar, then it can be obtained from the knowledge base and relevant information can be returned to the user.

Much of the research work is in progress for information extraction from other type of content such as image which is a part of image processing field. Also audio content are analyzed for information extraction.

VI. CONCLUSION

After analyzing different web mining techniques for extracting knowledge from web data for creating semantic web, it can be concluded that the unstructured data present on the web can be scanned to create ontologies to populate the knowledge base of the search engine. The information inserted in this knowledge base is in structured form that computers can understand. This information from the knowledge base can be used by the computers to better serve the web user's query. Thus we can add semantics to the current web by extracting knowledge and creating ontologies to create the semantic web.

REFERENCES

[1] Jayatilaka A.D.S, Wimalarathne G.D.S.P, “Knowledge Extraction for Semantic Web Using Web Mining”. The International Conference on Advances in ICT for Emerging Regions - ICTer2011

[2] Alani, Harith; Kim, Sanghee; Millard, David E.; Weal, Mark J.; Hall, Wendy; Lewis, Paul H. and Shadbolt, Nigel R. (2003). “Automatic ontology-based knowledge extraction from web documents”. IEEE Intelligent Systems, 18(1), pp. 14–21.

[3] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, “Web Mining – Concepts, Applications and Research Directions”.

[4] Tim Berners-Lee, James Hendler, Ora Lassila, “The Semantic Web”, available at: http://semanticweb.org/wiki/Semantic_Web

[5] Berners-Lee, Tim; James Hendler and Ora Lassila (May 17, 2001). "The Semantic Web". Scientific American Magazine.

[6] Wilson, Tracy V., "How Semantic Web Works". HowStuffWorks.com, available at: http://www.howstuffworks.com/semantic-web.htm

[7] Gruber, Tom (1993), "A Translation Approach to Portable Ontology Specifications".

[8] Marek Obitko, "What is Ontology", available at: http://www.obitko.com/tutorials/ontologies-semantic-web/what-is-ontology.html

[9] “Web Ontology Language”, available at: http://en.wikipedia.org/wiki/Web_Ontology_Language

[10] Michael K. Smith, Chris Welty, Deborah L. McGuinness, "OWL Web Ontology Language", available at: http://www.w3.org/TR/owl-guide/