Mining Web Data Based on Ontology and SWRL

Vhatkar Varsha Vilasrao and P. C. Bhaskar

***Abstract*—In our day to day work, if we could share knowledge across systems, costs would be reduced. However, because knowledge bases are typically constructed from scratch, each with their own idiosyncratic structure, sharing is difficult. Thus recent research has focused on the use of Ontologies to promote sharing. Ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base. If two knowledge bases are built on a common Ontology, knowledge can be more readily shared, since they share a common underlying structure. Ontologies provides, a shared and common understanding of a domain that can be communicated between people and application systems. Ontology is defined as partial specification of conceptual vocabulary used for formulating knowledge-level theories about a domain of discourse. Ontology is applied in domains like natural disaster management system, medicine, military intelligence, cooking, enterprise, jobs, agriculture, Wikipedia, automobiles and so on. Ontologies have become common place as a way to represent both knowledge and data. The knowledge provided by Ontology is extremely useful in defining the structure and scope for mining Web content, identify the meaning of data without labels and algorithm learns automatically rules expressed in Semantic Web Rule Language (SWRL) and this helps find semantic data. The Ontology is recommended for formal semantics and it is look like a backbone of every semantic web applications. In this paper a technique is proposed for ontology design using semantic web rule language. The proposed technique is useful for normal search as well as semantic search according to their senses, for semantic search the system uses the WordNet dictionary. In this system, Ontology is represented as a set of concepts and their inter-relationships relevant to some knowledge domain.**

***Index Terms*—Ontology, semantic web(SW), semantic web rule language(SWRL), support vector machine(SVM), world wide web consortium(W3C), web ontology language(OWL).**

I. Introduction

The rapid expansion of hugely unstructured data on the Web is causing several problems such as an increased difficulty of extracting potentially useful knowledge. Distilling relevant information from unstructured data, such as content from Web pages, can be both challenging and time consuming. What does it mean to mine data from the Web, as opposed to on other sources of information? The Web contains a mix of many different data types, and so in a sense subsumes text data mining, database data mining, image mining, and so on. The Web contains additional data types not available in large scale before, including hyperlinks and massive amounts of (indirect) user usage information. Spanning across all these data types there is the dimension of time, since data on the Web changes over time. There is data that is generated dynamically, in response to user input and programmatic scripts. If an operation is performed across all the data on the Web, then it must scale to the size of the Web. If not performed over the entire Web, is it Web data mining or just text data mining, database data mining, etc? One could argue that extracting information from Web pages, even a very small subset of all Web pages, is an instance of Web data mining, because the format and types of data that appear on the Web are different in general than other kinds of data. The World Wide Web also known as the Web, is a system of interlinked hypertext documents accessed via the Internet. With a web browser, one can view web pages that may contain text, images, videos, and other multimedia and navigate between them via hyperlinks. Also most of the data on web is weakly structured and largely unorganized. The user uses the search engine for information extraction, which can help people to search for required content in web pages. Most of the web search engines logically organize web pages into a structured, indexed semantic document for query of information. So some web engineering methodologies use the ontology in their development process. Ontology is a backbone of SW application. The design of new Ontologies during SW application development is having lack of focus. But web content is not always easy to use. Because of its unstructured and semi structured nature of web pages and the design of web sites. Therefore introduced the idea of semantic web which deals to the construction of an understandable semantic result over the existing web content so its support for better information processing and web services.

A. *Ontology*

Ontologies provide a shared understanding of a domain. They provide background knowledge to systems to automatize certain tasks. By the process of annotation, knowledge can be linked to Ontologies, for example: “Angelina Jolie” linked to concept Actress.

In our Ontology we also know that an actress always is female and a person. Ontologies allow the creation of annotations: machine-readable and machine-understandable content. If machines can understand content, they can also perform more meaningful and intelligent queries. Distinction of Jaguar the animal and the car. Combination of information that is distributed on the Web.

Ontology provides the means for describing explicitly the conceptualization behind the knowledge represented in a knowledge base. Ontologies are the backbone of the Semantic Web, They provide the knowledge that is required for semantic applications of all kinds,They are essential to

Manuscript received May 2, 2012; revised June 18, 2012.

Vhatkar Varsha Vilasrao is with Department of Technology, Shivaji University, Kolhapur, Maharashtra, India (varsha.vhatkar@gmail.com).

P. C. Bhaskar is with Department of Electronics, Shivaji University, Kolhapur, Maharashtra, India (pxbhaskar@yahoo.co.in).

DOI: 10.7763/IJIET.2012.V2.180

semantic annotation. An Ontology is an engineering artefact consisting of, A vocabulary used to describe a particular view of domain, An explicit specification of the intended meaning of the vocabulary. Often includes classification based information. It has constraints capturing background knowledge about the domain. Ideally, Ontology should capture a shared understanding of a domain of interest and provide a formal and machine manipulable model. An Ontology is an engineering artifact, it is constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary. Thus, an Ontology describes a formal specification of a certain domain, shared understanding of a domain of interest. So Ontology is an explicit specification of a conceptualisation. Ontologies typically have two distinct components:

- Names for important concepts in the domain: for example, elephant is a concept whose members are a kind of animal, Herbivore is a concept whose members are exactly those animals who eat only plants or parts of plants, Adult Elephant is a concept whose members are exactly those elephants whose age is greater than 20 years
- Background knowledge/constraints on the domain: for example, adult Elephants weigh at least 2,000 kg, All Elephants are either African Elephants or Indian Elephants, No individual can be both a Herbivore and a Carnivore

B. *Semantic Web*

Difficulties to find, present, access, or maintain available electronic information on the web, need for a data representation to enable software products (agents) to provide intelligent access to heterogeneous and distributed information. The solution is the Semantic Web, “The Semantic Web is a major research initiative of the World Wide Web Consortium (W3C) to create a metadata-rich Web of resources that can describe themselves not only by how they should be displayed (HTML) or syntactically (XML), but also by the meaning of the metadata”. The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. In semantic Web, a goal of the Web was that, if the interaction between person and hypertext could be so intuitive that the machine-readable information space gave an accurate representation of the state of people's thoughts, interactions, and work patterns, then machine analysis could become a very powerful management tool, seeing patterns in our work and facilitating our working together through the typical problems which beset the management of large organizations.

C. *Semantic Web Rule Language*

SWRL rules are written as antecedent, consequent pairs [1]. In SWRL terminology, the antecedent is referred to as the rule *body* and the consequent is referred to as the *head*. The head and body consist of a conjunction of one or more *atoms*. At present, SWRL does not support more complex logical combinations of atoms. SWRL rules are similar to rules in Prolog or DATALOG language. In fact, SWRL rules are similar to DATALOG rules with unary predicates for describing classes and data types, binary predicates for properties, and some built in n-ary predicates.

- Example of SWRL: Jeevan belongs to the Class ChildOfMarriedParents.
- Axioms: Jeevan Type Person, Jeevan hasParent Seeta, Jeevan hasParent Gopal, Gopal hasSpouse Seeta. In this example, the symmetric property hasSpouse connects Gopal and Seeta. That’s why Jeevan is Son of Gopal and Seeta.
- Protege Syntax: Person(?X), hasParent(?X,?Y), hasParent(?X,?Z), hasSpouse(?Y,?Z) ChildOfMarriedParents

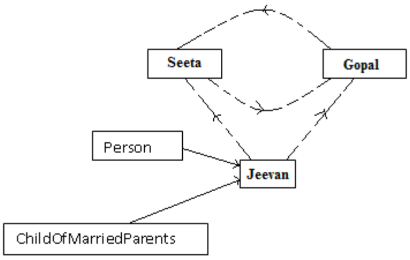

Fig. 1. SWRL example.

It can be described in functional syntax too:
Declaration (class(:ChildOfMarriedParents))
subClassof(:ChildOfMarriedParents:Person)
SWRLRule(Body(ClassAtom(:Person variable (var:X))
ObjectPropertyAtom(:hasParent variable(var:X) variable(var:Y))
ObjectPropertyAtom(:hasParent variable(var:X) variable(var:Z))
ObjectPropertyAtom(:hasSpouse variable(var:Y) variable(var:Z)))
Head(ClassAtom(:ChildOfMarriedParents variable (var:X))))

D. *Web Crawler*

Web Crawler is called as Web spiders or Robots, these are programs used to download documents from the internet. Simple crawlers can be used by individuals to copy an entire web site to their hard drive for local viewing. The work of crawler is easily parallelized, and dividing the URL space by domain seems like the best solution. A web crawler is a program that browses the World Wide Web in a methodical, automated manner or in a orderly fashion. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a web site, such as a checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from web pages, such as harvesting e-mail addresses usually for sending spam. In general, it starts with a list of URLs to visit, called the seeds. As the crawler visits there URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the Crawl frontier. URLs from the frontier are recursively visited according to a set of policies. The large volume implies that the crawler can only download a fraction of the web pages within a given time, so it needs to prioritize it downloads. The high rate of change implies that the pages might have already been updated or even deleted. A crawler always downloads just fraction of the web pages, it is highly desirable that the downloaded fraction contains the most relevant pages and not just a random sample of the web. This requires a metric of importance for prioritizing web pages. The importance of a page is a function of its intrinsic quality, its popularity in terms of links or visits, and even of its URL. The importance of a page for a crawler can also be expressed as a function of the similarity of a page to a given query.

II. Literature Overview

Manual approach, through which by observing a web page and its source code, the programmer can find some schemas from the web page, and write a program to identify, as well as to extract the data items. This approach is not suitable for a large number of pages. Wrapper induction [2], [3] a set of extraction rules are learnt from a set of manually labeled pages or data records. These rules used to extract data from similar pages. Automatic extraction [4] find data items have different roles in web pages, it is resolved at various levels: Semantic blocks, sections and data items, and several approaches are proposed to identify the mapping between data items having same role. Road Runner [5] works by comparing the HTML structure of two given sample pages belonging to a same page class, generating as a result schema for the data contained in the pages. From this Schema, some same pages can be extracted, but most of pages are heterogeneous, this method is more time consuming. Basically Ontology extract data from outside environment called context knowledge, infer or analyze the data, and then respond in real time to environmental situations by providing suitable services to user. As this is done in a context driven manner, the way context is represented rather important in developing such systems. The issues below are raised in [6]. Context knowledge is usually represented differently in various systems without a standard, causing poor interoperability, reusability, and distributed composition [7]. Further as the current representation is lack of semantics to fully represent the hierarchy of context knowledge, it is difficult to infer the knowledge. In domain Ontology OWL is used to define context knowledge the relationship in it and its property. On the other hand, SWRL is used to define inference rules directly using the terms defined in OWL. This integrates OWL knowledge and rules rather well. Protege [8] is used to develop OWL and SWRL. It provides graphical user interface for easy development and management of Ontology. According to [9] a rule axiom consists of an antecedent (body) and consequent (head), each of which consists of a (possibly empty) set of atoms. Atoms can be of the form c(x), p(x, y) same as (x, y) and different from (x, y), where c is an OWL description or data range, p is an OWL property, x and y are either variables, OWL individuals or OWL data values, as appropriate. Web classification [10] done by classification of web contents previously for this purpose SVM classifier was used. SWRL does not support negation atoms proposes an extension to OWL with general rules, namely E-SWRL, getting classical negation and default negation involved into SWRL rules. One cannot expect from the occasional user to bear with such long time intervals during his browsing and querying, nor is there any room for improvement. In [11] a complete policy-based management framework is presented, which includes a policy specification language and architecture for deploying policies, KAoS policy and domain services[12] use Ontology concepts encoded in OWL to build policies.

III. Motivation

The web pages, the data extracted from multiple data resources is different in the form of format and order. And lot of data stored on web pages, only present the data related to user query.

An existing technique that involve checking the similarity between a text and the seed list of words. This system is combination of Ontology and the semantic web. In this approach we use a domain Ontology according to that classifies the keywords into different categories. If user search for some word then the user also include the class and subclass for that word after that also synonyms of that word will be added and search those many words after that we will get all the information related to those words.

IV. Conclusion

Semantic Web conceptually large interlinked database, contents are formally defined and its utilization is maximum. It has reasoning capability. Semantic web is easy and efficient for information searching, accessing, extracting, interpreting and processing. Semantic Web has more accuracy, less semantic heterogeneity. It consists of content, formal semantics and presentation. Semantic Web has text simplification and clarification. This system structure Ontology with semantic web to facilitate search engines to support information sharing from the point of view of users and which deals with the entire universe of knowledge, and it shares the knowledge with other web application systems. This is used to share knowledge in order to obtain best result from search engine based on an Ontology and Semantic Web. This system combines both classification of things and semantic search. The term Semantic Web is often used more specifically to refer to the formats and technologies that enable it. Especially the collection, structuring and recovery of collected linked data are empowered by various web languages. Ontologies are the structural frameworks for organizing information and are used in artificial intelligence, the semantic web, system engineering, software engineering, biomedical informatics, library science, enterprise bookmarking and information architecture as a form of knowledge representation about the world or some part of it. Normally, Internet Search Engines employ many computers to index the Internet via web crawling. Such systems may allow for users to voluntarily offer their own computing and bandwidth resources towards crawling web pages. By spreading the load of these tasks across many computers, cost that would otherwise be spent on maintaining large computing clusters are avoided, in further work this can be implemented by distributed web crawling.

V. FLOW OF SYSTEM

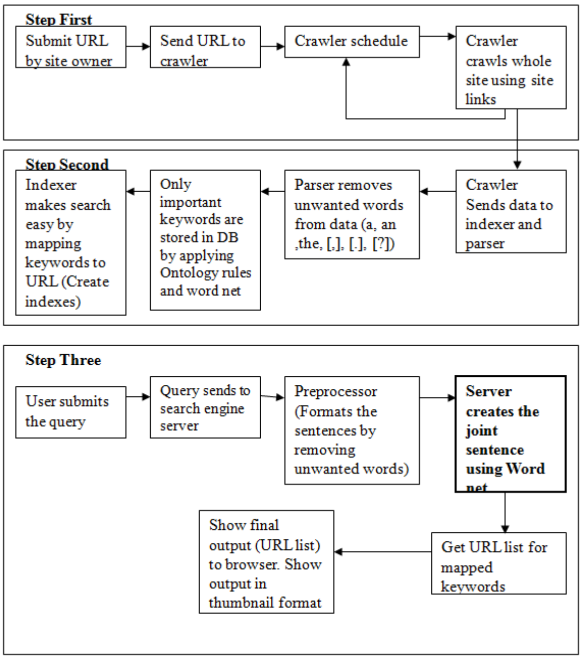
Step First
Submit URL by site owner
Send URL to crawler
Crawler schedule
Crawler crawls whole site using site links
Step Second
Indexer makes search easy by mapping keywords to URL (Create indexes)
Only important keywords are stored in DB by applying Ontology rules and word net
Parser removes unwanted words from data (a, an ,the, [,], [.], [?])
Crawler Sends data to indexer and parser
Step Three
User submits the query
Query sends to search engine server
Preprocessor (Formats the sentences by removing unwanted words)
Server creates the joint sentence using Word net
Show final output (URL list) to browser. Show output in thumbnail format
Get URL list for mapped keywords

Fig. 2. Step wise flow of System.

ACKNOWLEDGMENT

At the outset, I would like to express my sincere thanks to Prof. P. C. Bhaskar whose supervision, inspiration and valuable guidance, helped me a lot to complete this paper. His guidance proved to be the most valuable to overcome all the hurdles in the completion of this research. I am also grateful to Prof. (Dr.) R. K. Kamat, Reader, Department Of Electronics, Shivaji University, Kolhapur for that moral boosting. I would like to thank Department of Technology, Shivaji University, Kolhapur for supporting this research. And last but not least, this acknowledgement would be incomplete without rendering my sincere gratitude to express my deepest to my parents and husband who stood by me and supported me in the completion of this research.

REFERENCES

[1] Y. Sun, J. Zhang, W. Zhao, and Y. Tian, “Managing and Refining Rule Set for SWRL,” *IEEE* 2008.

[2] A. Laender, B. Ribeiro-Neto, A. da Silva, and J. Teixeira, “A Brief Survey of Web Data Extraction Tools,” *SIGMOD* Record, vol. 31, no. 2, pp. 84-93, 2002.

[3] E. O’Neil, B. Lavoie, and R. Bennett, “Trends in the Evolution of the Public Web, 1998-2002,” *D-Lib Magazine*, vol. 9, no. 4, 2003.

[4] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” *Scientific American*, vol. 284 no. 5, pp. 35-43, 2001.

[5] R. Kosala and H. Blockeel, “Web Mining Research: A Survey,” *SIGKDD Explorations*, vol. 2, no. 1, June 2000.

[6] T. Strang and C. L. Popien, “A context modling survey” *The 6th International conference on Ubiquitous Computing*, Sept. 2004.

[7] C. H. Liu, K. L. Chang, J. Y. Jason. Chen, and S. C. Hung, “Ontology-Based Context Representation and Reasoning Using OWL and SWRL,” *IEEE*, 2010.

[8] *The Protege Ontology Editor* [Online]. Available: http://www.prot ég éstanford.edu/.

[9] J. Mei, *Ontologies and Rules in the semantic Web*, 2001.

[10] A. Sun, E. Lim, and W. K. Ng, “Web Classification using Support Vector Machines,” *ACM Workshop on Web Information and Data Management (WIDM’02)*, 2002.

[11] N..Damianou, “N.Duley, E. Lupu, and M. S. Ponder,” *A language for specifying security and management policies for distributed systems*, 2000.

[12] J. Bradshaw and A. Uszok, *Representation and reasoning based policy and domain services in kaos and no-mads*, 2003.