ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

Web Data Extraction using Ontology and SWRL

Ms. Vhatkar Varsha Vilasrao

***Abstract* — Ontology is used for formal semantics and it is look like a backbone of every semantic web applications. In this system a technique is proposed for ontology design using semantic web rule language. By this technique, existing web applications design methods may easily be upgraded for semantic web applications. 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. 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.**

Keywords- Ontology, Semantic Web(SW), Semantic Web Rule Language(SWRL) ,World Wide Web Consortium (W3C), Web ontology language(OWL)

I. INTRODUCTION

In World Wide Web, the web contains a lot of information, and it is a distributed all kinds of documents including hypertext, text, PDF files, software, images, video and audio files. Also most of the data on web is weakly structured and largely unorganized. The user uses the search engine for information extraction, that 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.

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

A. Ontology

An Ontology deals with the question concerning what entities exist or can be said to exist and how such a entities can be grouped related within hierarchy and subdivided according to similarities and difference. Ontology in a computer science is a formal representation of knowledge as a set of concept within a domain and the relationships between these concepts. And also ontology is a specification of a conceptualization. A body of formally represented knowledge is a based on a conceptualization, and the objects, concepts and the other entities that are assumed to exist in some area of interests and the relationship that hold among them. A conceptualization is an abstract simplified view of the world that we wish to represent for some purpose. Every knowledge base, knowledge-based system or knowledge-level agent is committed to some conceptualization, explicitly or implicitly. Ontology is an explicit specification of conceptualization. Ontologies are part of the W3C standards stack for the semantic web in which they are used to specify standard conceptual vocabularies in which to exchange data among systems, provide services for answering queries, publish reusable knowledge bases and offer services to facilitate interoperability across multiple, heterogeneous systems and databases. The key role of Ontologies with respect to database systems is to specify a data modeling representation at a level of abstraction above specific database designs (logical or physical), so that data can be exported, translated, queried, and unified across independently developed systems and services. Historically, Ontologies arises out of the branch of philosophy known as metaphysics, which deals with the nature of reality of what exists. This fundamental branch is concerned with analyzing various types or mode of existence, often with special attention to the relation between particulars and universal, between intrinsic and extrinsic properties and between essence and existence. The traditional goal of Ontological inquiry in particular is to divide the world "at its joints", to discover those fundamental categories, or kinds, into which the worlds objects naturally fall. During the second half of the 20th century, philosophers extensively debated the possible methods or approaches to building Ontologies, without actually building any very elaborate Ontologies themselves.

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

In the 1980s, the AI community began to use the term Ontology to refer to both a theory of a modeled world and a component of knowledge systems. Some researchers, drawing inspiration from philosophical Ontologies viewed computational Ontology as a kind of applied philosophy. In the early 1990s, the widely cited web page and paper "To word Principals for the Design of Ontologies Used for Knowledge Sharing" is credited with a deliberate definition of Ontology as a technical term in computer science. Ontologies are often equated with taxonomic hierarchies of classes, class definitions, and the subsumption relation, but Ontologies need not be limited to these forms. In early years of the 21st century the interdisciplinary project of cognitive science has been bringing the two circles of scholars closer together. Ontology concern with the study of being or existence. In computer and information science, Ontology is a technical term denoting an artifact, that is designed for a purpose, which is to enable the modeling of knowledge about some domain, read or imagine. Ontology engineering is concerned with making representational choices that capture the relevant distinctions of a domain at the highest level of abstraction while still being as clear as possible about the meaning of terms. As in other forms of data modeling, there is knowledge and skill required. The heritage of computational Ontology in philosophical Ontology is a rich body of theory about how to make Ontological distinctions in a systematic and coherent manner. For example, many of the insights of "formal ontology" motivated by understanding "the real world" can be applied when building computational Ontologies for words of data. When Ontologies are encoded in standard formalisms, it is also possible to reuse large, previously designed Ontologies motivated by systematic accounts of human knowledge or language. Ontology is acquired from the query interfaces and query result pages of training web sites. Although the training texts contain abundant information in that field. For improve an accuracy and efficiency of extraction, the system will complement Ontology iteratively in the process of extracting new pages. There are two types of Ontologies, Domain Ontology and Upper Ontology. A Domain Ontology models a specific domain, which represents part of the world. And an Upper Ontology is a model of the common objects that are generally applicable across a wide range of Domain Ontologies. It associate with object description.

B. Semantic Web

Semantic search seeks to improve search accuracy by understanding searcher intent and the contextual meaning of terms.

There are two major forms of Search, Navigational and Research. In Navigational search, the user is using the search engine as a navigation tool to navigate to a particular intended document. Semantic search is not applicable to navigational searches. In Research search, the user provides the search engine with a phrase which is intended to denote an object about which the user is trying to gather/research information. There is no particular document which the user knows about that she/he is trying to get to. Rather, the user is trying to locate a number of documents which together will give her/him the information she/he is trying to find. Semantic search as a set of techniques for retrieving knowledge from richly structured data sources like ontologies as a found on the Semantic Web. Such technologies enable the formal articulation of domain knowledge at a high level of expressiveness and could enable the user to specify his intent in more detail at query time. In order to understand what a user is searching for, word sense disambiguation must occur. When a term is ambiguous, meaning it can have several meanings for example, if one considers the word "bark", which can be understood as "the sound of dog," "the skin of tree," or "a three masted sailing ship", the disambiguation process is started. In this most probable meaning is chosen from all those possible. Such process make use of other information present in a semantic analysis system and takes into account the meaning of other words present in the sentence and in the rest of the text. The determination of every meaning, in substance, influences the disambiguation of the others, until a situation of maximum plausibility and coherence is reached for the sentence. All the fundamental information for the disambiguation process, that is, all the knowledge used by the system, is represented in the form of semantic network, organized on a conceptual basis. The Semantic web is the roadmap of a "man made woven web of data" that facilitates machines to understand the semantics, or meaning, of information on the world wide. The concept of Semantic Web applies the methods beyond linear presentation of information and multi linear presentation of information to make use of hyper structures leading to entities of hypertext. The Semantic Web as "a web of data that can be processed directly and indirectly by machines". Semantic Web has more advantageous than current web. Current web does not have formal semantics of its contents. These are machine readable but not machine understandable. Current web more similar or look like a book with consisting multiple hyperlinked documents. The book have index of keywords but if the user use the keyword which is missing. It is difficult or inconvenient for user because no formal semantics of keywords in indexes.

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

This limitation is eliminated by Ontology in that data is given well-defined formal meanings, understandable by machines. So in Semantic Web formal semantics of data are available via Ontologies and completely accessible to semantic search engines. The Semantic Web has many other advantages in terms of information searching, accessing, extracting, interpreting and processing. The Semantic Web as originally envisioned is a system that enables machines to understand and respond to complex human requests based on their meaning. It has remained in a permanent state of improvement of its framework of standards. It is understood today that understanding by machines within the available repositories of information requires prior systematic structuring of the contents and strategic proliferation of tools for performing such structuring in a layered approach. There is no escaping the fact, the availability of understandable and intelligible information, even in automated processing, needs time for preparation. The main purpose of Semantic Web is driving the evolution of the current web by enabling users to find, share and combine information more easily. Humans are capable of using the web to carry out tasks such as finding the Irish word for "folder", reserving the library book, and searching for the lowest price for a DVD. However, machines cannot accomplish all of these tasks without human direction, because web pages are designed to be read by people not machines. The Semantic Web is a vision of information that can be readily interpreted by machines can perform more of the tedious work involved in finding, combining, and acting upon information on the web. The Semantic Web is regarded as an integrator across different content, information applications and systems. It also provides mechanisms for the realization of Enterprise Information Systems. Rapid growth in the volume of data on the web provides the impetus for researchers to focus on the creation and dissemination of innovative Semantic Web technologies to facilitate automated processing. Often the term 'Semantics', 'metadata', 'Ontologies' and 'Semantic web' are used inconsistently. In particular, these terms are used as everyday terminologies by researchers and practitioners, spanning a vast landscape of different fields, technologies, concepts and application areas. Furthermore, there is confusion with regard to the current status of the enabling technologies envisioned to realize the Semantic Web. The idea of Semantic Web, able to describe and associate meaning with data necessarily involves more than simple XHTML markup code. It is based on an assumption that in order for it to be possible for machines to accurately interpret web content, far more than the natural language ordered relationships involving letters and words is necessary as underlying infrastructure attendant to semantic issues. Additions to infrastructure to support semantic functionality include latent dynamic network models that can, under certain conditions, be 'trained' to appropriately 'learn' meaning based on order data, in the process 'learning' relationships with order.

C. Web Crawler

Web Crawlers also called Web spiders or Robots, 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. A web crawler is one type of software agent. 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. Web crawlers that attempt to download pages that are similar to each other are called focused crawler or topical crawlers. In focus crawling only able to predict the similarity of the text of a given page to the query before actually downloading the page. A possible predictor is the anchor text of links, this was the approach taken in a crawler of web.

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

To use the complete content of the pages already visited to infer the similarity between the driving query and the pages that have not have been visited yet. The performance of a focused crawling depends mostly on the richness of links in the specific topic being searched, and a focused crawling usually relies on a general web search engine for providing starting points. A crawler may only want to seek out HTML pages and avoid all other MIME types. In order to request only HTML resources, a crawler may make an HTTP HEAD request to determine a web resource's MIME type before requesting the entire resource with a GET request. To avoid making numerous HEAD request, a crawler may examine the URL and only request a resource if the URL ends with certain characters such as .html, .htm, .asp, .aspx, .php, .jsp, .jspx or a slash. This strategy may cause numerous HTML web resources to be unintentionally skipped. Some crawlers may also avoid requesting any resources that have a "?" in them in order to avoid spider traps that may cause the crawler to download an infinite number of URLs from a web site. This strategy is unreliable if the site uses URL rewriting to simplify its URLs. Crawlers usually perform some type of URL normalization in order to avoid crawling the same resource more than once. The term URL normalization, also called URL canonicalization, refers to the process of modifying and standardizing URL in a consistent manner. There are several types of normalization that may be performed including conversion of URLs to lowercase, removal of "." and ".." segments, and adding trailing slashes to the non empty path component.

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] 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 [3] 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 [4] 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 [9]. Context knowledge is usually represented differently in various systems without a standard, causing poor interoperability, reusability, and distributed composition [9]. Further as the current representation are 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 [10] is used to develop OWL and SWRL. It provides graphical user interface for easy development and management of ontology. According to [10], 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. SWRL does not support negation atoms [11] proposes an extension to OWL with general rules, namely E-SWRL, getting classical negation and default negation involved into SWRL rules. The usage of reasoning services seems to imply applications distributed in nature. Reasoning problems in OWL are solved optimally in 2-N NEXP time [12], for OWL 2 it is 3-NEXP [13]. 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 [14], a complete policy-based management framework is presented, which includes a policy specification language and architecture for deploying policies, KAoS policy and domain services [15] use ontology concepts encoded in OWL to build policies. All these approaches did not address knowledge exchange and sharing among all the stakeholders in the e-business scenario.

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 classify

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

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 the number of links that related to those words.

IV. CONCLUSION AND FUTURE WORK

Semantic Web conceptually large interlinked database, contents are formally defined and its utilization is maximum. It have 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 content, formal semantics and presentation. Semantic Web has text simplification and clarification. This system structure an Ontology with semantic web to facilitate search engines to support information sharing from the point of view of users and which is 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. The term ontology has its origin in philosophy and has been applied in many different ways. The word Ontology comes from Greek. Ontology means 'existence'. It consists of a set of types, properties and relationship types.

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.

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 work. His guidance proved to be the most valuable to overcome all the hurdles in the completion of this research work.

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 work.

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 work.

REFERENCES

[1] Laender, A., Ribeiro-Neto, B., da Silva, A. and Teixeira, J., "A Brief Survey of Web Data Extraction Tools," SIGMOD Record 31(2):84-93, 2002.

[2] O'Neil, E., Lavoie, B. F., and Bennett, R., "Trends in the Evolution of the Public Web, 1998-2002,", D-Lib Magazine 9(4), 2003.

[3] Berners-Lee, T., Hendler, J. and Lassila, O., "The Semantic Web," Scientific American. 284(5):35-43, 2001.

[4] Kosala, R., and Blockeel, H., "Web Mining Research: A Survey," SIGKDD Explorations, 2(1), June 2000.

[5] Sun, A., Lim, E.-P., and Ng, W. K., "Web Classification using Support Vector Machines," ACM Workshop on Web Information and Data Management (WIDM'02), 2002.

[6] Wen Z ang,Kerui Chen,Fan Zhang , "Mining Data Records based on Ontology Evolution for Deep Web," 2010 IEEE V7.

[7] Yunchuan Sun, Junsheng Zhang, Wei Zhao, Yingjie Tian, "Managing and Refining Rule Set for SWRL," 2008 IEEE.

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

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

[10] The Protege Ontology Editor [Online] :http://protégé.stanford.edu/.

[11] Jing Mei, Ontologies and Rules in the semantic Web, 2001.

[12] S.Tobies, "Complexity results and practical algorithms for logics in knowledge repreentation", RWTH-Aachen University, Germen, 2001.

[13] Y.Kazakov, "SRIQ and SROIQ are harder than SHOIQ",2008.

[14] N..Damianou, N.Duley, E. Lupu, and M. Sloman. Ponder,"A language for specifying security and management policies for distributed systems", 2000

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

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

V. SYSTEM ARCHITECTURE

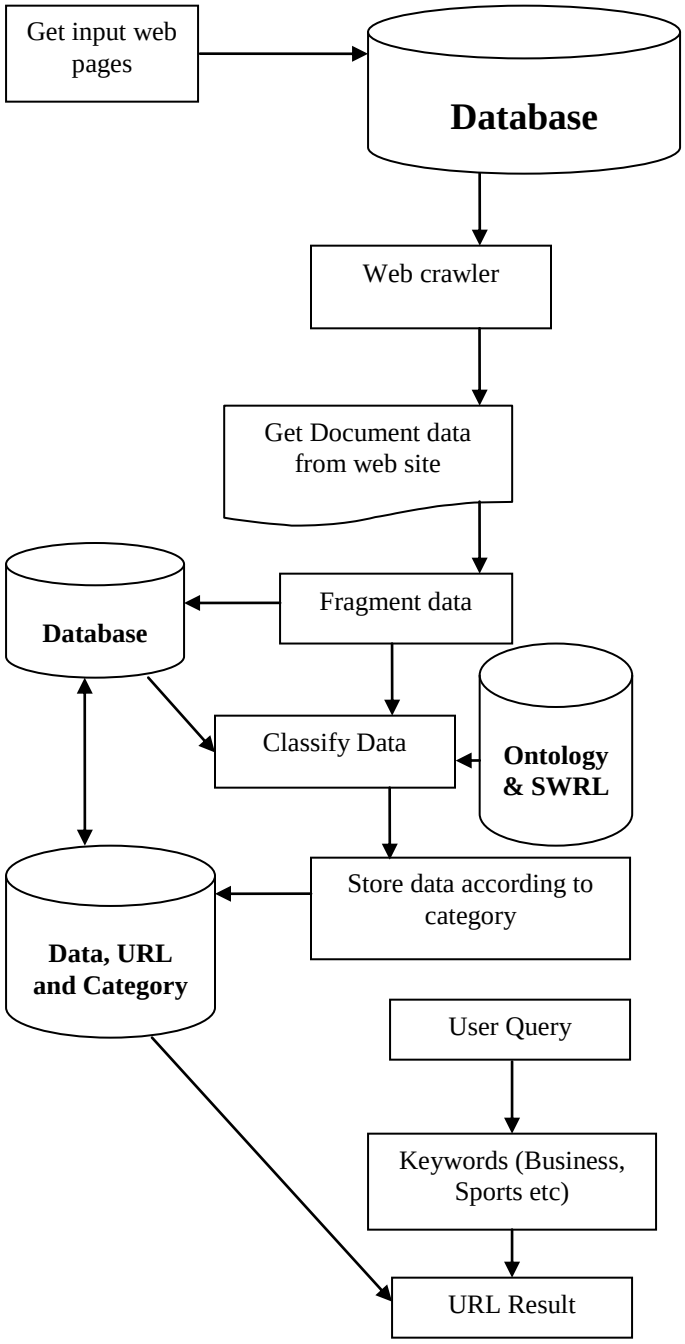

Fig. 1 Architecture of the system.

Personal Profile

Name : **Ms. Vhatkar Varsha Vilasrao**
Department : Department of Technology
University : Shivaji University
e-mail id : ***varsha.vhatkar@gmail.com***
City : Kolhapur
State : Maharashtra
Country : India.
Mobile No. : 09321607395.
Educational Qulification: M.Tech Computer Science and Technology(Pursuing, Shivaji University)

B.E. Computer Science and Engineering (Shivaji University)

Diploma in Industrial Electronics (Govt. Polytechnic Kolhapur)
Publication : Paper Present at,
"National Conference On Emerging Trends in Computer Science and Information Technology" NCETCSIT-2011
Participated in Dipex 2001 (Mumbai) For Project Presentation.

Ms. Vhatkar Varsha Vilasrao
Department of Technology,
Shivaji University,
varsha.vhatkar@gmail.com*, Kolhapur, India.*