International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 6 (2017), pp. 1397-1404

http://www.ripublication.com

Application for Automatic Annotation of Academic Institution Website

S.Sivaramakrishnan[1], Dr. T.Senthil Kumar[2], Dr. A.Valarmathi[3] & Mrs. S. Nalini[4]

[1]Student, Department of Computer Applications,

Head of the Department, Department of Computer Applications,

[3]Assistant Professor, Department of Computer Applications,

[1-4]University college of Engineering, BIT Campus, Anna University, Trichy-620 024, Tamilnadu, India.

Abstract

Semantic web, a proposed development of the World Wide Web in which data in web pages is structured and tagged in such a way that it can be read directly by computers. Annotation is a methodology for adding information to document at some level - a word, phrase or paragraph. This information is called "meta-data". The extensive pattern-matching notation of regular expressions enables to quickly parse large amounts of text to find specific character patterns; to validate text to ensure that it matches a predefined pattern (such as an e-mail address); to extract, edit, replace, or delete text substrings; and to add the extracted strings to a collection. Natural language processing is used to parse the data into various part of speech, so that tagging is done. Using regular expressions to extract content for the Institution website page and Producing Automated Annotation using Natural Language

Processing, and RDF meta-data for the HTML files. Automating the Ontology Engineering by performing activities programmatically and producing OWL based Ontology.

Keywords: Semantic, NLP, RDF, OWL, Linked, web, data, institution, annotation, automatic

I. INTRODUCTION

Some of the challenges for the Semantic Web include vastness, vagueness, uncertainty, inconsistency, and deceit. Automated reasoning systems will have to deal with all of these issues in order to deliver on the promise of the Semantic Web.

Vastness: The World Wide Web contains many billions of pages. The SNOMED CT medical terminology ontology alone contains 370,000 class names, and existing technology has not yet been able to eliminate all semantically duplicated terms. Any automated reasoning system will have to deal with truly huge inputs.

Vagueness: These are imprecise concepts like "young" or "tall". This arises from the vagueness of user queries, of concepts represented by content providers, of matching query terms to provider terms and of trying to combine different knowledge bases with overlapping but subtly different concepts. Fuzzy logic is the most common technique for dealing with vagueness.

Uncertainty: These are precise concepts with uncertain values. For example, a patient might present a set of symptoms that correspond to a number of different distinct diagnoses each with a different probability. Probabilistic reasoning techniques are generally employed to address uncertainty.

Inconsistency: These are logical contradictions that will inevitably arise during the development of large ontologies, and when ontologies from separate sources are combined. Deductive reasoning fails catastrophically when faced with inconsistency, because "anything follows from a contradiction". Defeasible reasoning and paraconsistent reasoning are two techniques that can be employed to deal with inconsistency.

Deceit: This is when the producer of the information is intentionally misleading the consumer of the information. Cryptography techniques are currently utilized to alleviate this threat. By providing a means

a) Components

The term "Semantic Web" is often used more specifically to refer to the formats and technologies that enable it. The collection, structuring and recovery of linked data are enabled by technologies that provide a formal description of concepts, terms, and relationships within a given knowledge domain. These technologies are specified as W3C standards and include:

Resource Description Framework (RDF), a general method for describing information RDF Schema (RDFS) Simple Knowledge Organization System (SKOS) SPARQL, an RDF query language Notation3 (N3), designed with human-readability in mind N-Triples, a format for storing and transmitting data Turtle (Terse RDF Triple Language) Web Ontology Language (OWL), a family of knowledge representation languages Rule Interchange Format (RIF), a framework of web rule language dialects supporting rule interchange on the Web.

II. LITERATURE SURVEY

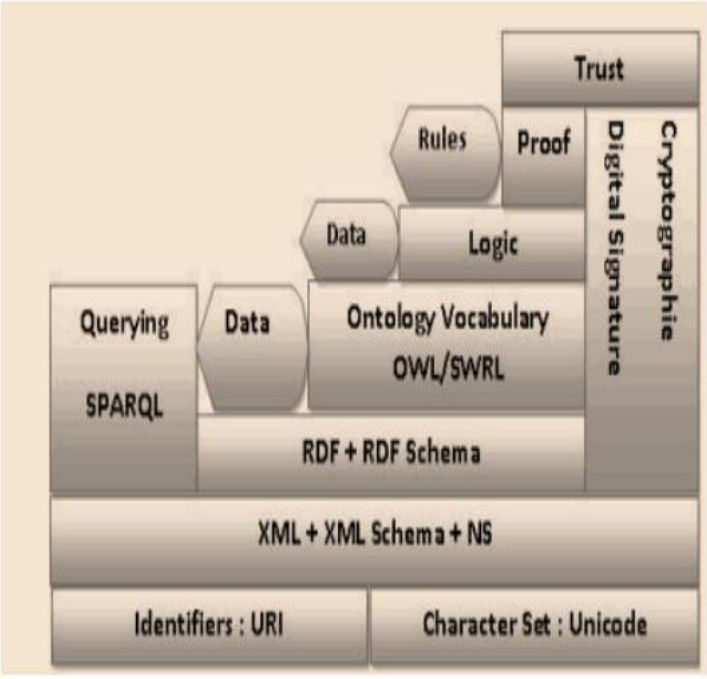

Figure 1. Architecture of Semantic Web

The bottom layers contain technologies providing common syntax. Uniform Resource Identifier (URI) provides means for uniquely identifying semantic web resources (entities)1 , while Unicode serves to represent and manipulate text in many languages

useful for exchanging symbols. The Extensible Markup Language (XML) is a markup language that enables creation of documents composed of structured data, and XML Schema allows the definition of grammars for valid XML documents. Semantic web gives meaning (semantics) to structured data. XML documents can refer to different namespaces to make explicit the context (and therefore meaning) of different tags. XML Namespaces provide a way to use markups from more sources. Semantic Web aims to connect data together, which needs to refer more sources in one document. The explained two layers are nowadays broadly accepted, and the number of XML documents is growing quickly. XML is the first step in the right direction, but it only formalizes the structure of a document and not its content. The Resource Description Framework (RDF)2 is a framework for creating statements in a form denoted by triples. This form enables the representation of information about resources in the form of graph and can be seen as the first layer where information becomes machine understandable: According to the W3C recommendation) , RDF "is a foundation for processing metadata; it provides interoperability between applications that exchange machine understandable information on the Web". The components of each RDF document consist of three types of entities: Resources (subjects and objects), properties (predicates/relations). Resources represent Web pages, parts or set of Web pages, or anything (real-world object) that can have a URI. Properties are specific attributes, or relations describing resources. The combination of a resource together with a property having a value for that resource forms a Statement (known as the subject, predicate and object). A value is either a literal, a resource, or other statement. A Statement which may be represented as a triple of the form (Subject, Property, Object) asserts that a resource recognized by the subject, has a property whose value recognized by the object (either another resource or a literal). Consequently, a property is a binary relationship between two resources or between a resource and a literal value.

RDF is basically a directed graph with labelled edges and partially labelled nodes. The definition of a simple modelling language on top of RDF is realized by the RDF Schema (RDFS)4 which includes classes, IS-a relationships between classes and properties, and properties characterized by domain/range restrictions. RDF and RDF Schema are structured in XML syntax, but they do not use the tree semantics of XML. An extension of RDFS including more advanced constructs to describe semantics of RDF statements based on description logic is provided by Web Ontology Language (OWL)[4].

At the layer of ontology vocabulary it is possible to query any RDF-based data (i.e., including statements involving RDFS and OWL) with the use of the latest RDF query language (SPARQL)

The remaining layers are Proof and trust. The top layers contain technologies that are not yet standardized which require the ability to check the validity of the statements

made in the (Semantic) Web, and Trust to derive statements will be supported by (a) verifying that the premises come from trusted sources and by (b) relying on formal logic during deriving new information.

Semantic annotation is the process that creates semantic labels of documents for the semantic Web, aiming to support advanced searching (based on concepts), reasoning about Web resources and the information visualization based on ontology. Additionally annotation is used to convert syntactic structures into knowledge structures. In other terms, semantic annotation consists to generate specific metadata and usage schema, enabling new information access methods and extending the existing ones.

Automatic semantic annotation can be realized on the base of automatic annotating algorithms: such as PANKOW (Pattern-based Annotation through Knowledge On the Web) and C-PANKOW (Context-driven and Pattern based Annotation through Knowledge on the Web)

III. EXISTING SYSTEM

The current institutional website is purely syntactic. The institutional website contains keywords and these keywords are used for producing the website while searched and looked for. Content based relation and linking is not available for data within the website.

The standard followed by current website is web2.0.

IV. PROPOSED SYSTEM

This application helps to create RDF files for the HTML files and produce meta-data for linking the data within the website.

It contains modules to tag the parts-of-speech using natural language processing (NLP) library, that contains list of words to perform parts-of-speech separation, which helps to identify noun.

Then the RDF files are generated using the xml namespace available in vb.net.

The RDF files are interlinked through data present with in tags.

The RDF graph visualization is performed for interlinked data.

V. ARCHITECTURE

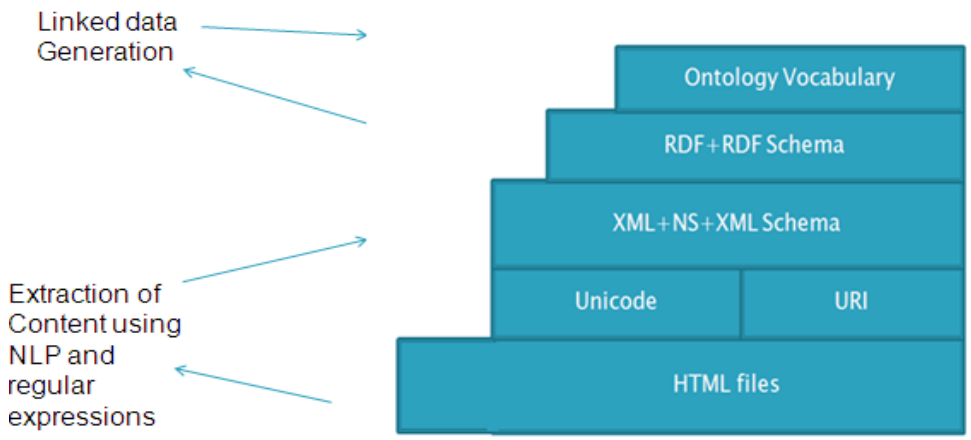

Figure 2. Proposed System Architecture

a) MODULES

Data Extractor-Extract data using regular expression patterns

Noun Finder-Part-of-speech identifier

RDF creator-RDF file generation

Linker-RDF linked data Generation & Linked data Visualization

Ontology generator-Ontology Engineering

VI. UML DIAGRAMS

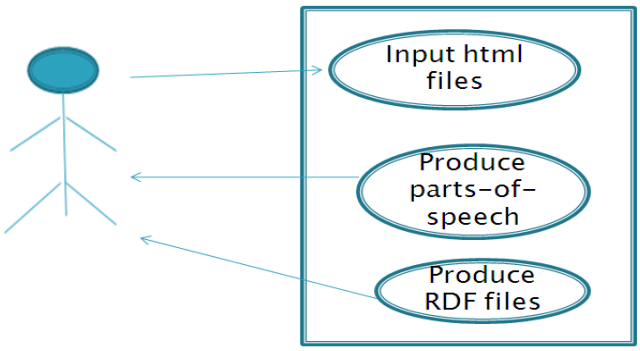

Figure 3. Use-case Diagram

VII. RESULTS

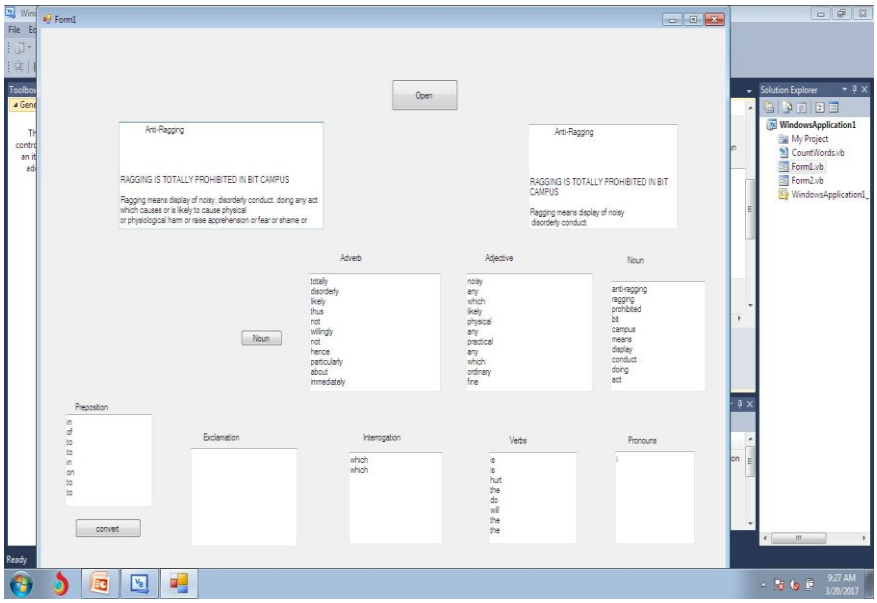

Figure 4. Noun Finder

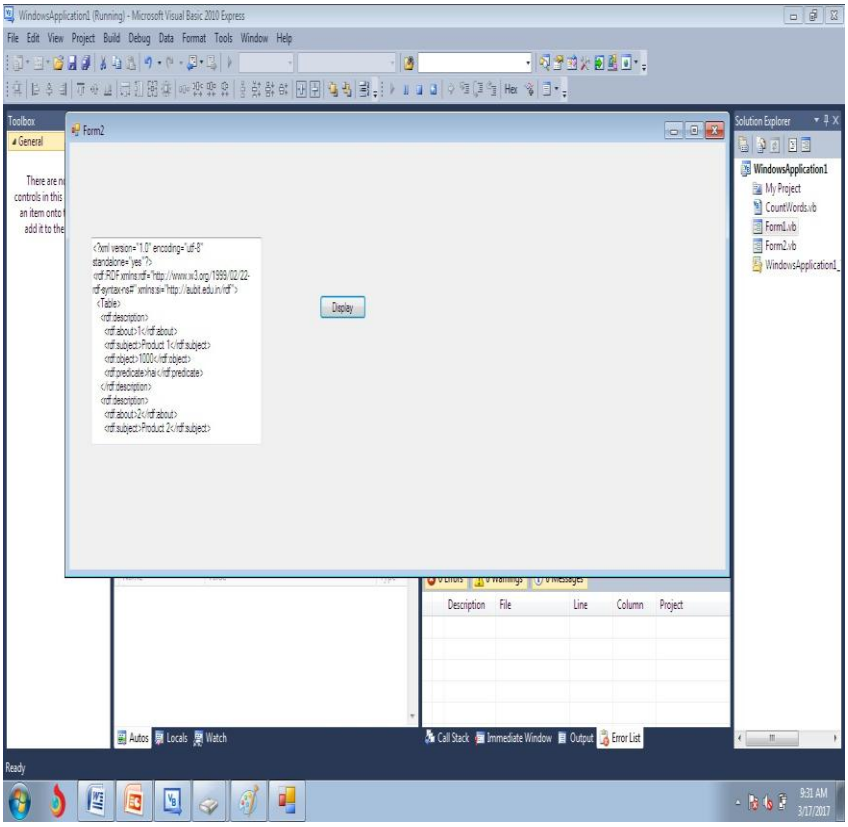

Figure 5. RDF Creator

VIII. CONCLUSION AND FUTURE WORK

The institutional website is currently working with keyword based approach.

The meta keyword and crawling based meta data are used for search engine optimization.

The institutional website is upgraded to semantic version from existing web 2.0 syntactic version, helps in better search results in search engines.

The RDF metadata helps to produce linked data.

Ontology based classification is provided for the institutional website.

REFERENCES

[1] Ontologies for Software Engineering: Past, Present and Future M. P. S. Bhatia, Akshi Kumar and Rohit Beniwal Indian Journal of Science and Technology, *Vol 9(9), DOI: 10.17485/ijst/2016/v9i9/71384, March 2016*

[2] *Knowledge Extraction and Modeling* from Scientific Publications by Francesco Ronzano and Horacio Saggion,2016

[3] *Semantic Annotation: The Mainstay of Semantic Web* by Thabet Slimani,2010

[4] *Ontology Issues and Applications Guest Editors' Introduction* by Fred Freitas, Heiner Stuckenschmidt, Natalya F. Noy,2007.

[5] *Ontology Engineering: An Application Perspective by* Salim K. Semy, Kevin N. Hetherington-Young, Steven E. Frey ,2004.