SEMANTIC INTEGRATION OF HETEROGENEOUS WEB DATA FOR TOURISM DOMAIN USING ONTOLOGY BASED RESOURCE DESCRIPTION LANGUAGE

Mrs. Jayaprabha P
Professor,
Department of Computer Applications,
Vidyaa Vikas College of Engineering & Technology,
Tiruchengode, Tamilnadu, India – 673 601.
Email: mppjaya@yahoo.co.in

Dr. Saradha A
Professor,
Department of Computer Science Engineering,
Institute of Road & Transport Technology,
Erode, Tamilnadu, India – 673 601.
Email: saradha_irtt@yahoo.com

ABSTRACT

World Wide Web (WWW) becomes the biggest repository of information and the most of the data in the web are in unstructured text format. Search engines returns inappropriate result to the users query and the machines find it difficult to integrate the information available in the World Wide Web due to unstructured information. During information retrieval from internet, the search engines take care of the locations instead of caring about semantics of the information. The above shortcomings are overcome with the evolution of the next generation web "semantic web". Semantic web maintains the web in a structured form and makes web accessible data more amenable to machine processing. It becomes a challenging task to the research community to convert, integrate voluminous of legacy text data that are unstructured and semistructured, into semantic web format. This paper proposes a generic approach to transform the existing web contents related to tourism domain into semantic web format using RDF with ontology. Such semantic web will be advantageous for semantic search and it provides better interoperability.

Keywords- Semantic Web, Ontology, OWL, Resource Description Framework (RDF), Semantic interoperability.

1. INTRODUCTION

The World Wide Web is the biggest repository of information ever assembled by man. It contains documents related to all fields, but maximum data in the internet are in unstructured text form. This unstructured information is only readable by humans and not by machines. Search engines find it difficult to retrieve relevant contents.

When user needs data related to tourism, the search engines searches data throughout web repository and provides some content that are relevant to the search topic and some contents are irrelevant to search topic and sometimes fails to provide data, because the existing web content are scattered, unstructured, document oriented, closed system and are not integrated. This makes the user unsatisfactory. In semantic web, the web contents are carefully designed with metadata. The web contents are universally accessible interlinking of content based on imputed semantics such as concepts, definitions or structured argumentation. The Semantic web contents are decentralized, heterogeneous and open. The semantic web has no central data repository, no central agreement on meaning and no central policy on terminology or structure. When user searches data, it provides most relevant data and the user gets satisfactory.

For e.g. to locate any information in the internet, it takes considerable amount of human hours and further more user want to perform series of task, in order to locate an information about tourist places in Chennai, hotel, shopping mall etc., user has to visit serried of web pages for integration of content and reasoning about them. This can be overridden by semantic web.

The challenge of the semantic web therefore is to provide a language that expresses rules for the data to be exported on to the web. RDF is an ontology based language used to implement semantic web.

This paper centers around tourism domain, how to convert conventional heterogeneous tourism data into Semantic web format RDF.

2. RELATED WORK

Semantic web development technologies are used for converting existing heterogeneous unstructured sports data into structured RDF using tools such as ANNIE, GATE[9].

Current web applications are mostly database driven, developers design a database schema and then construct the application logic on top of the schema. These applications are centralized and rely on their own relational database, limiting the possibilities for data integration. To overcome these shortcoming "A flexible integration framework for semantic web 2.0 applications" has been developed [2].

Rhizome is an experimental, open source content management framework used to capture and represent informal, human-authored content in a semantically rich manner. Rhizome aims to help bring about a new kind of commons—one of ideas. This commons wouldn't comprise just a web of

interlinked pages of content, as is the current World Wide Web, but a web of relationships between the underlying ideas and distinctions that the content implies: a permanent, universally accessible interlinking of content based on imputed semantics such as concepts, definitions, or structured augmentation.

Building Finder is a running application that integrates satellite imagery, geospatial data and structured and semistructured data from various online data sources using Semantic Web technologies. Users can query an integrated view of these sources and request Building Finder to accurately superimpose buildings and streets obtained from various sources on satellite imagery. The data sources integrated by Building Finder are heterogeneous not only in terms of the data, but also in terms of how the application accesses the sources [4].

Many research communities have exploited semantic web technologies to build interoperable applications. Consider multimedia databases-many multimedia documents such as images, video and sound records reside in huge databases of production companies, museums and TV channels. For these documents to be available in semantically rich manner, they must be automatically processed and annotated with the aid of knowledge representation languages. But processing and describing multimedia documents involves a lot of uncertain information.

For uncertainty resulting from inability to precisely define concepts and assign degrees of truth [1]. Fuzzy Reasoning Engine has been constructed to overcome the above shortcomings.

The semantic web development technologies are used only creating new web content and not for existing web content. It is proposed to convert existing tourism data into semantic web format.

3. WEB DEVELOPMENT TECHNOLOGIES

Web contents may be stored in either structured form or unstructured form. The unstructured web contents are developed using HTML, XML etc whereas structured web contents are developed using specialized tools and techniques.

3.1 Unstructured Information

Unstructured information's are only readable by humans and it is not interoperatable. These informations are written in any of the following language.

- HTML
- DHTML
- XML etc.,

3.1.1 HTML (Hyper Text Markup Language)

HTML is a language for creating web pages in the internet. Most of the web contents are mainly written in HTML.

Technology:

Markup means that contain sequence of characters in documents indicating the role document contents. It takes the form of words between tag.

e.g. <B> Chennai</B>

HTML is divided into two parts namely head and body. Head contains the title of the program and body contains actual information.

Drawbacks

HTML is mainly used for representation of content and not for interpretation of data or content. Here the tags are predefined. Web pages developed using HTML are static in nature. It is used for client side programming.

3.1.2 DHTML (Dynamic HTML)

Technology:

Dynamic HTML is a collection of technologies used together to create interactive and animated web pages by using a combination of a static markup language (such as HTML), a client-side scripting language (such as JavaScript), a presentation definition language (such as CSS), and the Document Object Model. DHTML allows scripting languages to change variables in a web page's definition language.

Drawbacks

DHTML are difficult to develop and debug, due to varying degrees of support among web browsers of the technologies involved.

3.1.3XML (Extensible Markup Language)

In XML, the tags are user defined and it is mainly used for interpretation. It is uniform data-exchange format.

XML has brought great features and promising prospects to the development of the Semantic Web. Using XML, one can describe document types for various domains and purposes.

Technology

Currently, there are numerous techniques and tools available for XML, e.g., SAX (Simple API for XML), DOM (Document Object Model), XSL (Extensible Style sheet Language), XSLT (XSL Transformation), XPath, XLink. XPointer and XML parsers are available in different languages and for different platforms. Using XML, one can describe document types for various domains and purposes.

Drawbacks

XML tags are user-defined and not unique, so that information are not machine operatable.

3.2 Structured Information

Information written using ontology based language is called structured information that

overcomes the drawbacks of unstructured information.

RDF is a framework for developing semantic web. It is language for expressing information in machine processable form.

Semantic web

Semantic web is an idea coined by Tim Berners-Lee, the creator of the WWW. Semantic Web is often described as a web for machines as opposed to a web to be read by humans. It allows users to organize and browse the Web in ways more suitable to the problems they have at hand.

These machine-interpretable descriptions allow more intelligent software systems to be written, automating the analysis and exploitation of web-based information. Software agents will be able to create automatically new services from already published services, with potentially huge implications for models of e-Business.

Semantic web uses many ontology languages to describe semantic data. Some of the ontology languages are as follows

- RDF (Resource Description Framework)
- OWL (Web Ontology Language)
- DAML (DARPA Agent Markup Language)
- SPARQL (Simple Protocol and RDF Query Language)
- GRDLL (Gleaning Resource Descriptions from Dialects of Languages)
- OIL (Ontology Inference Layer)

Ontology

An ontology is an explicit specification of the concepts in a domain and the relations among them, which provides a formal vocabulary for information exchange.

Information integration from different sources needs to be a shared by understanding of the relevant domain. Knowledge representation formalisms provide structures for organizing this knowledge, but provide no mechanisms for sharing it. Ontologies provide a common vocabulary to support sharing and reuse of knowledge.

3.2.1 Resource Description Framework

RDF provides a means for adding semantics to a document without making any assumptions about the structure of the document. It adds Meta information to Web documents.

The Resource Description Framework attempts to address XML's semantic limitations. It presents a simple model that can be used to represent any kind of data. This data model consists of nodes connected by labeled arcs, where the nodes represent web resources and the arcs represent properties of these resources.

The data model of RDF provides three object types: resources, property types and statements.

- **A *resource*** is an entity that can be referred to by a address at the WWW (i.e., by an URI). Resources are the elements that are described by RDF statements.
- **A *property*** defines a binary relation between resources and/or atomic values provided by primitive data type definitions in XML.
- **A *statement*** specifies for a resource a value for a property. That is, statements provide the actual characterizations of the Web documents.

Every statement has 3 parts:
- Subject
- Predicate
- Object

Example:- "Delhi is the capital of India http://india.ac.in/delhi.

This sentence has the following parts:
- Subject (Resource) http://india.ac.in/delhi
- Predicate (Property) Capital
- Object (Literal) "India"

This can be also be visualized as a graph- where subjects and objects are graph nodes and the predicates define directed arcs from a subject to an object.

Figure.1 RDF Graph

3.2.2 Introduction to OWL

OWL is a language for describing web information in structured format. OWL stands for Web Ontology Language designed to be interpreted by computers to analyze and interoperate, built on top of RDF and embedded in XML. It has three sublanguages namely

- OWL Lite
- OWL DL (includes OWL Lite)
- OWL Full (includes OWL DL)

OWL is a part of the "Semantic Web Vision" They are

- Web information has exact meaning.
- Web information can be processed by computers.

Owl Designed for Processing Information

OWL was designed to provide a common way to process the content of web information (instead of displaying it). OWL was designed to be read by computer applications (instead of humans).

OWL is different from RDF

OWL and RDF are much of the same thing, but OWL is a stronger language with greater machine interpretability than RDF. OWL comes with a larger vocabulary and stronger syntax than RDF.

3.2.2.1 Protégé OWL

The Protégé-OWL editor is an extension of Protégé that supports the Web Ontology Language (OWL). OWL is the most recent development in standard ontology languages, endorsed by the World Wide Web Consortium to promote the Semantic Web Vision.

Tourism Application created using Protégé OWL

Important properties of protégé are

- OWL Class Specification
- Properties
- Forms
- Individuals
- Metadata

The OWLClasses view can be used to edit hierarchies of concepts. Details of the selected class are shown in the right part of the screen. The upper part of this area allows users to add comments, labels and other annotations. The lower part displays logical characteristics of the selected class.

Protégé-OWL is seamlessly integrated with classification tools. These tools can be used to reveal inconsistencies and relationships between classes and individuals. The results of the classification are displayed on the OWLClasses tab, and can be easily navigated and analyzed.

The Properties tab can be used to edit characteristics of properties in the model.

The Individuals tab can be used to acquire instance data. The forms shown in the right half of the screenshot are generated automatically from the class definition. For example, if a class has a property of type single xsd:string, then the system would automatically display a text field widget for entering strings.

Protégé-OWL can also be used to edit RDF Schema models. The user interface will adjust to the selected language profile and display simpler widgets for rdfs:Classes.

The Protégé community has already contributed numerous extensions of the base platform. Among the most popular of these extensions is OWLViz, which can be used to visualize OWL ontologies graphically.

4. PROPOSED SYSTEM

The proposed system is the conversion of existing unstructured tourism data into structured web format RDF.

4.1 Information Extraction from Unstructured Text

There are plenty of web sites providing various data of tourism information on the web, including information about sights, routes, hotels, restaurants, tickets and so on, but almost none of them has the comprehensive travel information alone. If a tourist wants to search enough information to plan a trip, he has to constantly switch back and forth among a lot of various web sites. If the tourism data from different sources can be integrated into a whole, concentrative semantic data can absolutely make the query more convenient and the abundant semantic information behind the integrated data can also make the query more accurate and make data sharing and reusing more easily.

Data integration is the process of combining data residing at different sources and providing the user with a unified view of data. Data form different sources often have different perspectives, so they may overlap with each other. For example, “Madras” and “Chennai” refer to the same city, but they have totally different labels. Again the travel route of “One-day tourism in Beijing” contains the sight of “The Great wall”, while these two instances have no shared property.

Instance matching can resolve the semantic conflicts among heterogeneous data sources and find the connotative relation between instances extracted from different sources.

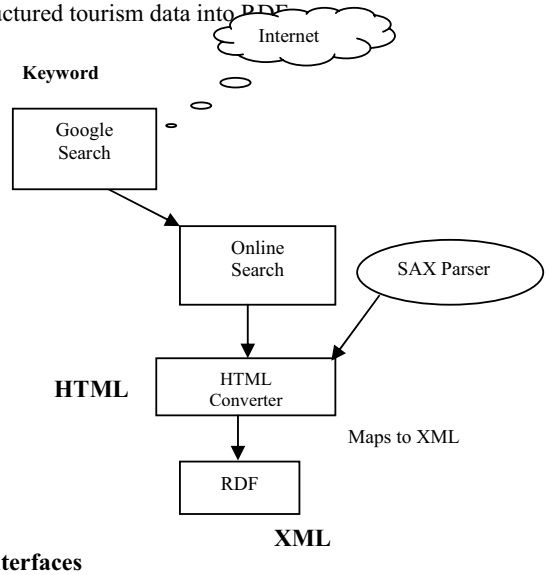
4.2 Domain of Text:

Domain selected for conversion of unstructured data into RDF is tourism. The popularity of the WWW resulted a flurry of websites covering tourism related information covering almost everything in the universe.

Tourism is one of important domain referring to many factors and has plenty of domain knowledge, which is the essential base of travel information systems. But the technologies are available only for creating new tourism based web content in structured semantic web format and not for converting existing unstructured web content into semantic web format – RDF.

5. BLOCK DIAGRAM

There are three modules for converting unstructured tourism data into RDF



5.1 Interfaces

5.1.1 Downloading unstructured text using online search

This module extracts heterogeneous tourism related text based on the keyword we specify in the online search dialogue. This module links Google search engine and extracts all the related travel information and stores in the specified path and in specified file name. The search result may be either single url with single

web page or thousands of url with more web pages. Now these documents are in HTML format. For example, use keyword tour in the online search module and this module downloads all the web pages related to keyword tour and stores in the name of Tourism1, Tourism2, Tourism3 etc., This is shown in the fig. 2.

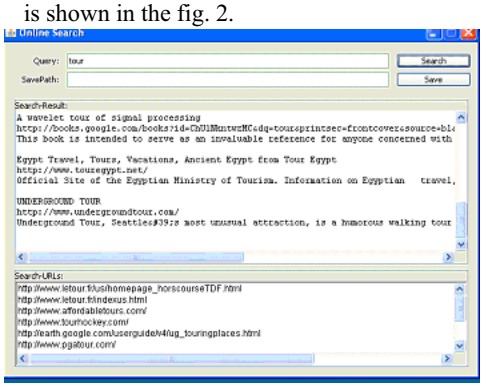


Figure 2. Downloading url's

From the downloaded url's, consider the first url ie Tourism1.html and it is shown in the fig.3.

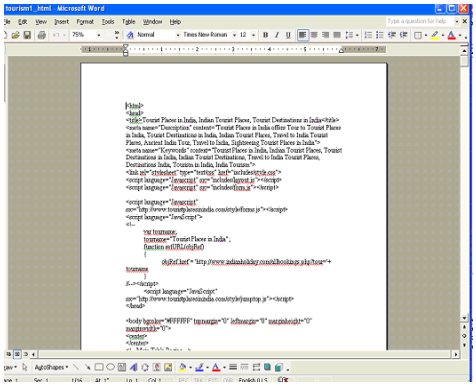


Figure 3. Tourism1.html

5.1.2 Converting HTML into XML

Tidy is software which automatically converts HTML into XML. It has some restriction that all the opening tags in HTML should have close tag and it should not contain any comment line etc. So this can't be used for the above implementation. So we used a module called HtmlConvertor which in turn uses SAX parser to read HTML data and converts to XML. It reads each and every tag in HTML and converts into related XML tag. While reading HTML tag, all the opening tags should have closing tag, otherwise it will show error in particular line, then we have to rectify this error manually. In this way HTML content are converted into XML.

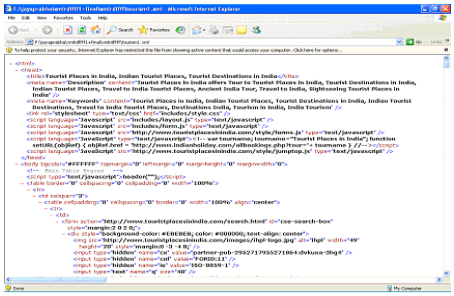


Figure 4. Tourism1.xml

Now this Tourism1.html is converted into Tourism1.xml by HtmlConvertor module and shown in the fig. 4

5.1.3 Mapping Process

Travel2.owl is an ontology based RDF file which contains the travel related information in turns of classes, instances, sub instances, object property, relation, data property, general axioms etc., This file has more 500 classes namely travel, destination, contact, accommodation, hotel, sight seeing etc.,

Accommodation has object property hasRating , isOfferedAt etc., and sub instances are collection in turn, TwoStarRating, OneStarRating, ThreeStarRating etc.

Destination has data property has Activity.Activity is disjoint with Relaxation, Sightseeing, Sports etc., Relaxation is sub class of Yoga.

Now the xml tags are mapped with classes in Travel2.owl and converted into RDF and it is shown in the fig. 5.

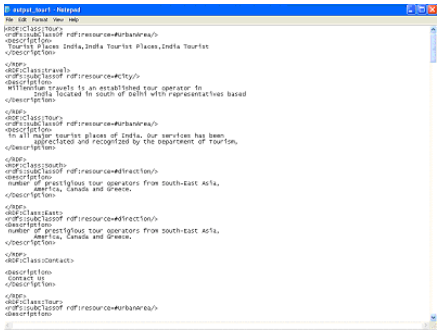


Figure 5. Tourism1.rdf

Contact has object property hasEmail, hasStreet, hasZipcode etc., XML tag is mapped with classes, sub classes, object property, data property etc., in travel2.owl. If mapping occurs the corresponding is converted into RDF and stored in output file. Now this RDF file will be available for semantic search.

6. Sample HTML into XML, XML into RDF

6.1 HTML

```
<html>
<head>
<title>Tourist Places in India, Indian Tourist Places, Tourist Destinations in India</title>
</head>
</html>
```

6.2 XML

<html>
<head>
<title>Tourist Places in India, Indian Tourist Places, Tourist Destinations in India</title>
</head>
</html>

6.3 RDF

<RDF:Class:Tour>
<rdfs:subClassOf rdf:resource=#UrbanArea/>
<Description>
Tourist Places India,India Tourist Places,India Tourist
</Description>

7. CONCLUSION

Travel related heterogeneous unstructured data from various web sites is extracted and using tools, it is converted into semantic web format called RDF. Implementation has been done in java. This data can be made available for web using semantic search engine.

REFERENCES

1. Madche Alexander and Steffen Staab. 2001. Applying Semantic Web Technologies for Tourism Information Systems. Research Paper.
2. "Owl web ontology language guide," Tech. Rep., W3C Recommendation, February 2004, available at http://www.w3.org/TR/owl-guide.
3. Werthner Hannes and Francesco Ricci. 2004. E-commerce and tourism. Communications of the ACM 47(12):101-105. New York: ACM Press.
4. Steffen Stub, " Uncertainty and the semantic web", University of Kolenz-Landau, Published by IEEE computer Society in 2006.
5. Eyal Oren, Armin Haller, Manfred Hauswirt, Benjamin Heitmann and Stefan deckar, "A flecible integration framework for semantic web 2.0 applications" published by IEEE computer society in 2007.
6. Yufei Li, Yuan wang and Xiaotao Huang, "A relation-based search engine in semantic web", IEEE transactions on knowledge and data engineering, vol. 19, no.2 feb 2007..
7. Martin Michalowski, José Luis Ambite, Snehal Thakkar, and Rattapoom Tuchinda, "Retrieving and semantically integrating heterogeneous data from web", published by IEEE Computer Society.
8. Downes S. (2005). Semantic networks and social networks, in The Learning Organization Journal. Finin T., Ding L. and Zou L. (2005).
9. "Standardization unstructured textual data into semantic web format" by Abdul Nazeer and Abdul Haleem , Department of Computer Engineering National Institute of Technology, Calicut.
10. ONTOURISM: SEMANTIC ETOURISM PORTAL byYing Ding, Institute of Computer Science, University of Innsbruck Innsbruck, Austria.

BIOGRAPHY

Jayaprabha received her Post Graduate Degree in Master of Computer Applications (MCA) from the University of Bharathidasan, Trichy in 2002 and Master of Philosophy in Computer Science from the Bharathiyar University, Coimbatore in 2004. She is a Assistant Professor and Head in the Department of MCA, VVCET (affiliated to Anna University of Technology, Coimbatore). She is pursuing PhD under Anna University of Technology, Coimbatore and her main research areas are Semantic web, Web mining & Ontology engineering.