Semantic Web Development for Tourism Domain from Conventional Web and Improving the Semantic Search by Providing Different Methods of Categorization

Jayaprabha P
Department of Computer Applications, VidyaaVikas College of Engineering & Technology, Tiruchengode, Tamilnadu, India – 673 601.

Saradha A
Department of Computer Science Engineering, Institute of Road & Transport Technology, Erode, Tamilnadu, India – 673 601.

ABSTRACT
World Wide Web is the biggest repository that contains all sorts of information. User can search necessary information from the web. All the information in the WWW are unstructured and human readable. Search engine provides result based on the user query is not satisfactory due to unstructured and semi-structured nature of web document. The advent of semantic web overcomes this difficulty, the information in this web are structured, machine readable and it is very easy to integrate. Now-a-days, most of the websites are developed based on the semantics, but there is no technology to convert existing web information into RDF. The proposed approach converts existing web information related to tourism domain in to semantic web using Resource description Framework. To obtain more relevant result the converted RDF web source are categorized and stored in the repository based on different algorithms.

Keywords

1. INTRODUCTION
The Web is becoming a dominant information repository. However, retrieval of relevant information is still a challenging task for most of its users. Ambiguity of words is one of the main hindrances in information retrieval[4][5]. Employment of semantic technologies in search systems is seen as a promising approach to improve the current state of the art [6]. Semantic technologies are applied in different ways: semantic annotations of content [7]; clustering of retrieved documents according to topics[8]; powerful querying languages [9]; or creating structured semantic models of retrieved documents[10].

Information present in the traditional web are unstructured, semi-structured, syntactic nature, difficult to integrate and more human readable. Search engine faces difficulty to extract information from traditional search. Drawbacks of this search engines has been discussed in following section. Data in the Semantic Web is defined and linked in a way that can be used for more effective discovery, automation, integration and reuse across applications. This data can be shared and processed by automated tools as well as people.

Difference between traditional and semantic search engine
At the core, a semantic search engine has the ability to understand the relationships between keywords, phrases or parts of speech within a search phrase, therefore allowing it to understand the underlying meaning of the entire phrase. For example, a semantic search engine would be able to easily distinguish the differences between the following phrases made up of the same ‘keywords’ but with obvious different implications:
- Chennai
- Madras

So how is semantic search different from traditional web search?
In the example above, the phrases are made up of the different keywords, while the subject/action relationships are same. In traditional web search, which are based on ranking algorithms, since the relationships between the sentence parts are unknown, the engines would return identical or nearly identical results, even though it was being asked two completely different questions. Additional problems with web search also arise when the keywords are too specific, producing few or no results, or too general, in which case the results are overwhelming and irrelevant.

Alternatively, since semantic search technology understands the meaning of the above sentences, it would be able to produce highly relevant answers to the questions. The goal of semantics is to always provide the direct insights and answers needed to complete research tasks, rather than burying those ideas among scores of irrelevant documents.

The final purpose of this engine is enhancing performance of traditional search engines (especially Precision and Recall). It’s possible through understanding the context of documents and queries. One of the most important parts of this type is annotator which responsible for generating metadata for crawled pages. It is needed to generate some metadata for user’s query in order to detect its context. Here usually after traditional retrieval, combine matching RDF graphs to obtain better quality of results.

16
2. RELATED WORK:
The next generation of Tourism Information System is expected to be: enable semantic-based information processing, exhibit natural language capabilities, facilitate inter-organization exchange of information in a seamless way, and evolve proactively in tandem with dynamic user requirements [2], [3].

The OntoWebbersystem[11] is an ontology-based approach to website management. It facilitates the design, creation, generation and maintenance of Web sites using a set of software tools. It also enables the personalization of Web site views based on individual users. Another notable approach is the Hera project[12], which is a methodology that supports the design and engineering of Web Information Systems (WIS) using Semantic Web technology. The main focus of the Hera project is to support Web design and implementation particularly hypermedia aspects.

3.0 TECHNOLOGIES USED
3.1 Semantic web
The Semantic Web is used for representing information in the World Wide Web in a machine-readable format, such that it can be used by machines not just for display purposes, but for automation, integration and reuse across applications.

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.

In the words of Tim Berners-Lee, “The Semantic Web is a web of data, in some ways like a global database,” and the Semantic Web effort is developing "languages for expressing information in a machine processable form." The Semantic Web makes use of structured text, rather than natural language, to identify knowledge and its relationship with other knowledge or data. Berners-Lee's original vision involves the action of intelligent software "agents" on computers and handheld devices that would act autonomously to both retrieve data and interact with Web sites through specialized tagging of data and content on these sites.

The Semantic Web functions because of a highly specialized type of data and information tagging that can be implanted within Web pages.

Fundamentally, each Semantic Web tag, known as a triplet, links together a subject, verb, and object, creating a relationship between them. Three simple examples of a semantic tag might be: <Boston>is in the state of> <Massachusetts>; <Beacon Hill> is a neighborhood of> <Boston>; <I> like <Boston>. These two statements each consist of two nouns separated by a verb and are interlinked with one another through Boston.

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

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

3.2 Ontology
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. Ontology is a fundamental component for achieving the Semantic Web. Ontology has the capability to solve a number of problems in tourism. This includes: 1) enabling interoperability of heterogeneous platforms; 2) standardization of business models, business processes, and knowledge architectures; and 3) serving as a model of knowledge representation for the generation of knowledge-based information services [1].

3.2 The Resource Description Framework
RDF provides a means for adding semantics to a document without making any assumptions about the structure of the document. It is an XML application customized for adding 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. It should be noted that this model is essentially a semantic network, although unlike many semantic networks, it does not provide inheritance.

The nodes/arcs model also means that RDF is inherently binary. However, this does not restrict the expressivity of the language because any n-array relation can represented as a sequence of binary relations. RDF can be exchanged using an XML serialization syntax.

The basic syntax consists of a Description element which contains a set of property elements. The about attribute identifies which resource is described. The property rdf:type is used to express that a resource is a member of a given class, and is equivalent to the instance-of link used in many semantic nets and frame systems. There are a number of abbreviated variations of the RDF syntax, which is an advantage for content providers but requires more complex RDF parsers.

It is important to note that all of these syntaxes have a well-defined

```xml
<?xml version="1.0"?>
<RDF xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:xml="http://schema.org/general#">
<Description about="http://www.state.edu/users/jsmith"/>
</RDF>
```

An RDF Instance.

```xml
<?xml version="1.0"?>
<RDF xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:xml="http://schema.org/general#"
xmlns:u="http://schema.org/university#">
<u:Chair about="http://www.state.edu/users/jsmith"
g:name="Jane Smith"/>
</RDF>
```

An Abbreviated RDF Instance. Mapping into the RDF data model, and thus avoid some of the problems with
representational choices in basic XML. Nevertheless, it is still easy to create different representations for a concept.

To prevent accidental name clashes between different vocabularies, RDF assigns a separate XML namespace to each vocabulary (these vocabularies, called schemas, can be formally defined using RDF Schema as discussed below).

This approach has two disadvantages. First, since namespaces can be used with any element and RDF schemas need not be formally specified, it is possible to write RDF statements such that it is ambiguous as to whether certain tags are RDF or interspersed tags from another namespace. Second, namespaces are not transitive, which means that each RDF section must explicitly specify the namespace for every schema that is referenced in that section, even for schemas that were extended by a schema whose namespace has already been specified.

A significant weakness of RDF is that it does not specify a schema inclusion feature. Although namespaces allow a document to reference terms defined in other documents, it is unclear as to whether the definitions of these terms should be included. In fact, it is unclear what constitutes the definition of a term.

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 a resource a value for a property. That is, statements provide the actual characterizations of the Web documents.

The Semantic Web is a web of data. There is lots of data we all use every day, and it’s not part of the web.

4. ABOUT TOURISM

Tourism plays major role in entertaining all sorts of people from young age to elder, poor person to rich one’s. One reason for high information exchange ratio among different players in tourism industry is the tourism product itself. In comparison to many other products, which are sold online, tourism product is immaterial, heterogeneous and non-persistent. Each of these characteristics has influence on the information exchange within tourism industry.

A trip usually includes many parts such accommodation, transportation, insurance, visa services, guide services, excursions in the destination. Due to the heterogeneity of the travel product, travel agency consultant or a person who is planning the trip itself must have access to different sources of information.

Tourism product is also immaterial, meaning that traveler cannot see or touch the tourism product before the trip. That is why reliable information about destination, accommodation options and other parts of the tourism product is extremely important for both people working in tourism industry and tourists themselves.

Tourism product cannot be stored in storage. If a hotel room or a seat in an airplane remains empty today, this is lost revenue for the tourism company. This is a reason why effective distribution and inventory management are key factors in the tourism business.

5. PROPOSED WORK

Domain selected for proposed work 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.

The proposed work is to convert existing unstructured heterogeneous tourism data into semantic web format such as Resource Description Framework and the same semantic web resources are categorized using two different techniques to enrich the search result of the user query. The work is categorized into six modules namely

1. Preprocess
2. Process1
3. Process2
4. Search

Each module will be elaborated in the following section.

5.1 Preprocess

This module extracts heterogeneous tourism related text based on the keyword and converts into semantic web format RDF. This module contains three sub modules namely

1. Downloading unstructured text using online search
2. Converting HTML into XML
3. Mapping Process

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. 1.

Fig. 1 Downloading url's
From the downloaded url’s, consider the first url Tourism1.html and it is shown in the fig. 2.

**Fig. 2 Tourism1.html**

5.1.2 Converting HTML into XML

Tidy is a 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 HtmlConverter 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.

Now this Tourism1.html is converted into Tourism1.xml by HtmlConverter module and shown in the fig. 3.

**Fig. 3 Tourism1.xml**

5.1.3 Mapping Process

Travel2.owl is an ontology based RDF file which contains the travel related information in terms 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, sightseeing etc., Accommodation has object property hasRating, isOfferedAt etc., and sub instances are Collection in turn, TwoStarRating, OneStarRating, ThreeStarRating etc.

Destination has data property hasActivity.

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. 4.

**Fig. 4. 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.

5.2 Process1 – Categorize offspring

After converting the existing web content into semantic web, now these web contents are categorized using four different process. As the World Wide Web changes dynamically, it’s impossible to structure a perfect ontology including the relations of all the web resources. In Semantic web, a class is a type of web resources. A class has some properties and it can be a subclass of another class. Thus, it cannot be matched by the ontology-based search engines for a web resource which is not defined as a class in the structured ontologies. To use Web resources efficiently, it becomes an important and emerging issue to detect the potential relations on existing ontologies.

For this purpose, a web resource categorization method is used to extract potential relations among the Web resources. The process1 module extracts the relation by using the knowledge such as the class hierarchy in ontologies, the relations of word in dictionaries, and the description of Web resources.

In this Process1, it is intend to categorize the classes well defined in ontologies. This is a traditional method in Semantic Web research area. If a class has a relation to a category, its subclasses (“rdf:subClassOf”) could have a weaker relation to
the category. The equivalent classes (“owl:equivalentClass” and “owl:sameAs”) have equivalent relation strength.

**Process 1** is performed by using a recursive function RER(c0, a, rc0,a). Let c0 be a class in a category a with the relation strength as rc0,a. If class c, which is an equivalent class or a subclass of class c0 in category a, has not been categorized to category a, RER(c0, a, rc0,a) can categorize class c to category a with relation strength as rc0,a. By recursively calling itself, RER(c0, a, rc0,a) categorizes all the offspring classes of c0. Here, given a coefficient k for the degressive strength of relation in the hierarchical structure. While k = 1 equivalentClass—relation

\[ rc,a = k \times rc0,a \]

Eq.(1)

\[ W_{i,j} = Wi,j \times \log(Ndf) \]

Eq.(3)

\[ tfi,j = \text{number of occurrences of } i \text{ in } j \]

\[ dfi = \text{number of categories containing } i \]

\[ N = \text{total number of categories in } A \]

Since Wi,j is largely affected by the number of classes in a category, we normalize Wi,j in [0, 1], and \( W_{i,j} \) denotes the normalized value which can be calculated by Eq. (4).

Consider the class names having \( W_{i,j} \) larger than a threshold minW as the important class names for the category. In the preliminary experiments, we got good result of the categorization while giving a threshold minW ranging from 0.02 and 0.04. Here, we specified it as 0.03. By considering both effectiveness and efficiency, the algorithm is described concretely as follows. For each important class name cName, we extract all the classes having class name cNamein category a from all the classes by function getClass(cName). From these classes, extend category a with each class c which denotes a class having a class name of cNamebut not existing in category a. For a set of classes having class name cNameand existing in category a, getR(cName, a) is a function to get the relation strengths of the classes in the set to category a, and avg(getR(cName, a)) denotes the average. Then, the relation strength of c to a can be calculated by Eq. (5).

\[ rc,a = \text{avg(getR(cName, a))} \times W_{i,j} \]

Eq.(5)
In the Process2, extend the categories with all the offspring classes of the newly categorized classes by executing Process1 again.

Fig.6 Process2 – Categorize by important class name

In the above process, the semantic web resources are categorized for efficient search of data for user query.

5.3 Search:
This module is used to find precision and recall factor for particular keyword that occurred in all the above four process. That is relation strength of a category and the classes.

Fig.7 Search module to find relation strength among four process
6. CONCLUSION AND FUTURE ENHANCEMENTS
Tourism related heterogeneous unstructured data from various web sites is extracted and using tools, it is converted into semantic web format called RDF. The proposed work is to convert existing unstructured heterogeneous tourism data into semantic web format such as Resource Description Framework and the same semantic web resources are categorized using four different techniques to enrich the search result of the user query. Implementation has been done in java. In future this work will be extended to several domains.

7. REFERENCES