Dynamic Tourism Information System Using the Semantic Web

Fatima Faiza Ahmed [*1], Fatima Saima hmed [2]

Department of Computer Systems Engineering, Dawood University of Engineering & Technology, Karachi Pakistan

Allmed Laboratories, Karachi Pakistan

[*1]baloch_fatimafaiza@hotmail.com; [2]fatimasaima_ahmed@hotmail.com;

Abstract- **The semantic web is one of the most promising and emerging areas of computer science. Its ability to make information that is only human understandable to be understood and processed by machines increases its relevance and time efficiency. In this paper we present a semantic web application that contains a dynamic knowledge base holding hotel related information, which is updated at every run from the information provided in the hotels' websites, and a natural language querying platform where the user can enter his free form natural language queries. DL Query is used as the engine that performs the retrieval of data from the ontology. A simple user experiment shows that the system is time effective and helpful in decision making.**

Keywords- Information Extraction; Semantic Web; DL Queries; Hotel Search Model

I. INTRODUCTION

The area of tourism is one that already uses all the available internet technologies very extensively. The area is ver y dynamic because it is one of the world's largest industries and its growth shows an increase every year. It has been predicted by the World Tourism Organization [2] that by 2020 the tourists travelling around the globe will have increased by more than

200%. Almost all of the travelers rely entirely upon the internet to search out information about their destination, airfares, hotels, areas for sightseeing or shopping etcetera. In [3] it is shown that about 95% of Web users use the Internet to gather travel related information and about 93% indicate that they visited tourism Web sites when planning for vacations. This increase in usage is what has brought about this great rise in tourism information systems present on the World Wide Web, offering structured tourism related information. However, this information is normally isolated from each other and rarely includes details.

One can not predict the nature or background of the traveller or his internet search skills. Nevertheless, he has to search each requirement separately, open several web pages, compare information, and then draw his conclusions. The burden of finding data and connecting it to relevant information is placed entirely upon the user. The semantic web is a solution when it comes to reliable searching since it adds meaning to data that the web contains and becomes machine processable, enabling precise and in-depth searching. The application presented in this paper is a dynamic portal for the hotel selection parameter of tourism, based on semantic web standards. Queried information is extracted from the actual websites of the hotels and stored in the dynamic ontology, where from the information is formatted and presented to the user.

II. LITERATURE REVIEW

Because of the dynamic nature of the tourism industry, and the capability of the semantic web to effectively deal with all its requirements, various semantic web applications related to tourism have been researched upon and developed. In their paper titled "Applying Semantic Web Technologies for Tourism Information Systems" published in 2002, Alexander Maedche and Steffen Staab explained the relevance and applicability of the semantic web for tourism purposes. In 2003, Helmut Berger, Michael Dittenbach, and Dieter Merkl presented their research in "Querying Tourism Information Systems in Natural Language", in which they explained in detail how the Natural Language Processing took place.

The system I present here studies the hotel selection parameter in detail, and how the natural language is processed in user queries.

III. DYNAMIC TOURISM INFORMATION SYSTEM

A Tourism Information System (TIS) is a new business application that supports e-tourism and e-portals for different areas of tourism. It serves e-travel organizations such as hotels, travel agencies, car-rentals, leisure and environmental agencies and travel related information sources. In [1] we have presented an application that acts as an e-portal for hotel related information. The research presented was of a static Knowledge Base holding hotel related data, a Natural Language

DOI:10.5963/IJCSAI0303005

querying platform and SPARQL, the engine that performed the retrieval of data. Here we extend this research by making this e-portal dynamic, by extracting real-time information from the hotels' websites. The architecture of the proposed model is shown in Figure 1.

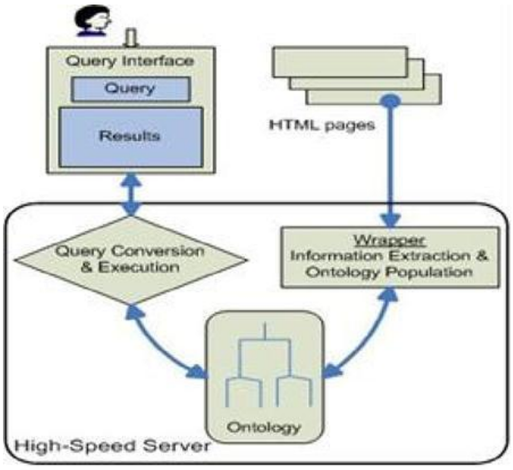

Fig. 1 Architecture of the SWEPT System

Using the current web, a tourist planning to travel has to visit multiple sites, compare a lot of information, spend hours searching for the correct facts, registering their personal information more than once.

The semantic web is the new generation web that facilitates meaningful search. This has enabled us to create a single platform where information from different tourism websites can be displayed, depending on the user queries.

IV. SOLUTION FOR VARIOUS E-TOURISM ISSUES

The importance of Tourism Information Systems is understood by the heavy demand and the need of the consumers. Hence, developers and producers of TIS spend a lot of time and effort in order to improve service for the ease and convenience of the consumers. Nevertheless, certain problems have always existed in creating a truly dynamic TIS.

After studying many different Tourism Information Systems, we found that there is a lack of standards in the tourism domain. There is no set format for the representation of different kinds of data. For example, some web sites represent prices and fees in US dollars, while others in Pounds, Euros, Rupees or Yen. Some web sites represent temperature in Centigrade, others in Fahrenheit. Some websites represent time in minutes, some in hours, and others in minutes and hours [3].

Our objective is to create a system that understands different ways to represent the same data, extract the relevant information and structure it into one format.

V. ARCHITECTURE

The architecture of the system is designed in a way to overcome the difficulties normally faced in building such systems. One of the greatest challenges is to meaningfully integrate information in different web sites, as they do not have any stand ard or set criteria to express transportation facilities, hotel and accommodation, weather conditions or leisure activities.

A. Hotel Search Model

One of the most important decisions to be taken before planning any journey is of a place to stay. A large number of daily searches are those of tourists intending to travel, and finding information on their destination. A user's decision on a hotel is based on many parameters. Sometimes trips are planned for leisure, sometimes hotel suitability is assessed based on tour plan and sometimes a hotel is selected for business meetings and gatherings by considering many factors like the number of guests and expenses. A hotel is selected after the analysis of a lot of factors. A user has to visit multiple sites and compare different attributes of different inns, like the features they offer and their rates per night. A benefit of providing all the comparative information on a single interface based on user selected parameters will provide a better and more efficient approach to hotel selection [9].The user is provided with an interface to enter his free form natural language queries. The query is interpreted at the server and converted to its equivalent DL query to fetch the data from the ontology.

B. Dynmic Ontology Design

The ontology we have termed as MyHotel Ontology incorporates information of various Pakistani hotels, their rates per

DOI:10.5963/IJCSAI0303005

night, facilities, and room availability. A graphical representation of the classes and their object properties is shown in Fig 2.

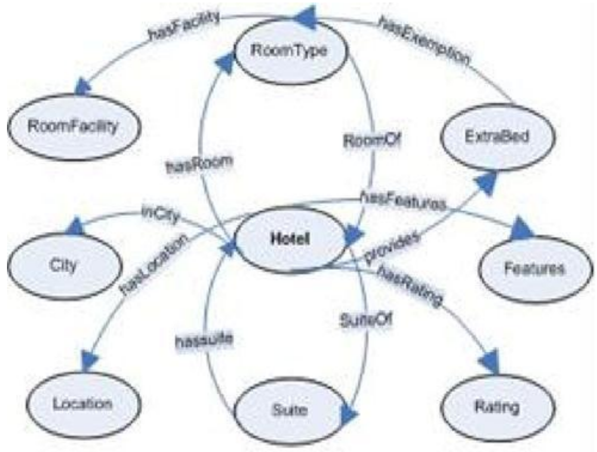

Fig 2. Classes and Object Properties of the MyHotel Ontology

Some of the characteristics of the ontology design are:

1) Task Centered Terms Selection:

An ontology is always constructed with a certain task in mind. This task focus restricts the content and structure of the ontology [5]. The MyHotel ontology is designed for querying purposes. The terms are selected from various Pakistani hotel websites in English.

2) Hotel as Center of Ontology:

After getting all relationships in place, we observe that the class "Hotel" is a center concept point in the ontology. This facilitates the incorporation of hotels in queries where it is required but not mentioned by the user. For example, if we consider the query "Is there any family room in Peshawer for less than Rs. 5000?" In this query, the term „hotel‟ is not used, but a human can easily understand that the query is asking about a hotel that offers a family room within the mentioned price.

3) Synonym Enrichment:

The rdfs: labels hold the information of the synonyms and acronyms of a concept initialized in our system. For example the term amenity is used in some sites instead of facility. An initialization module contains a camel case splitter method that populates the rdfs: label with the space separated terms to facilitate the terms matching for concept mapping. eg. RoomRatePerNight is stored in rdfs: label as "room rate per night". In this way a query containing different labels would be recognized so as to have the same results.

4) Individuals Selection:

We have defined rooms as Individuals in our ontology. We have done this because of two reasons. The first reason is the variation in the name of the room. For example, deluxe room in one hotel is termed as executive in other with different room facilities. Second important aspect is the rate associated with a room type. If one individual, for example 'StandardRoom' is fixed for all standard room types, the rate associated with it cannot be changed according to hotels.

5) *Information Extraction*:

To populate the ontology with online hotel data, a wrapper is needed that performs extraction of the data and instances related to selected ontological concepts from the website and then populate the ontology.

Currently, the wrapper developed extracts the room information from the trusted hotel websites that are selected and set in the system. No broker websites serving hotel information are selected. The selection is also made according to the consistency of the structure in representing room information. The wrapper identifies the table header and data records of html tables to extract the room type, breakfast (if available in the table), price information and availability using the attribute -value pair idea discussed in [10]. Figure 3 shows an RDF Snippet containing room instance with its rate.

```
<RoomType rdf:ID="PremierDeluxe4Y">
    <hasRatePerNight rdf:datatype="&xsd;float">330.0</hasRatePerNight>
    <roomOf rdf:resource="#Hotel_Y"/>
</RoomType>
```

Fig 3. RDF Snippet containing room instance with rate

DOI:10.5963/IJCSAI0303005

The information is extracted from different websites by parsing the html of the web pages, and extracted from the tables. We have started by extracting room types, rates per night and availability. For this purpose, we use HTML:: TableExtract which is a subclass of HTML:: Parser, which extracts information stored in a table within an html document.

Currently there are four ways to extract information using the HTML:: TableExtract. For our system, we have used IE by specifying Headers. This is the most flexible and adaptive technique for IE, involving the specification of the text in an array that you expect to appear above the tables of interest. After all headers have been matched in a row of that table, all cells following the columns that matched your headers are extracted. All other columns are ignored. Other than that, TableExtract automatically rearranges each row in the order of the headers provided.

C. Natural Language Query Processing

The results of any query are fetched by first converting the natural language query into its equivalent DL Query. For this purpose, we use our updated SWEPT (Semantic Web E Portal for Tourism) [1] system. An analysis of query logs of other major search engines has shown that, in practice, only 9% of the queries contain Boolean operators or the modifiers [9]. Most of them require that a query term must or must not be present in the searched pages. One cannot even be sure if a query is treated case sensitive or not. These issues do not arise in SWEPT, where Boolean operators, quantifiers and Uppercase or lowercase letters can be used. For example the query "Room with rates less than Rs 5000 per night in Karachi " contains a quantifier restriction "less than" with a quantifier value 5000. The conversion process by SWEPT for a hotel information query is as follows:

1) Sentence Splitting:

The natural language queries are split into words and multiple spaces are eliminated. All words are converted into their singular form, and predefined stop words such as of, on, for, in, to, list, show me, I want to know, How, When, Where, What etc. are eliminated.

2) Resource Tagging:

Then, the quantifiers such as “less than”, “more than”, “maximum”, “minimum” are identified and converted into operators like >, <, <= . The key concepts are then selected from the query and compared with the resource list in order to tag them with their respective ontology resource. Six types of tagging entities are used in SWEPT, which are Class (C), Individual (I), Object Property (PO), Datatype Property (PD), Filter Expression (FO) and Literal (L). Literal is a string stored for string datatype. For example, "Abul Hassan Isphani Road" is stored as a string for datatype property has Road, which if found in the user query is mapped to the corresponding datatype property. If we consider the user query: “room rate per night < 5000 in Karachi with AC” is tagged as: RoomType (class), ratePerNight (datatype property),

Karachi (Individual, with class “Location”), AC (Individual, with class roomFacility).

3) Graph Construction:

The classes are parsed to determine the possible shortest path between them. The shortest path connected graph constructed for the classes of the query given above is shown in bold with classes (vertices) and object properties (edges) in Figure 5. From that graph, it can easily be observed that the term 'hotel' is not mentioned in the query but due to its central location, it appears in the connected graph. The incorporation of the missing concept 'hotel' in the query graph has provided the real interpretation to the user query which can now be interpreted as: "hotels room rate per night < 5000 in Karachi AC".

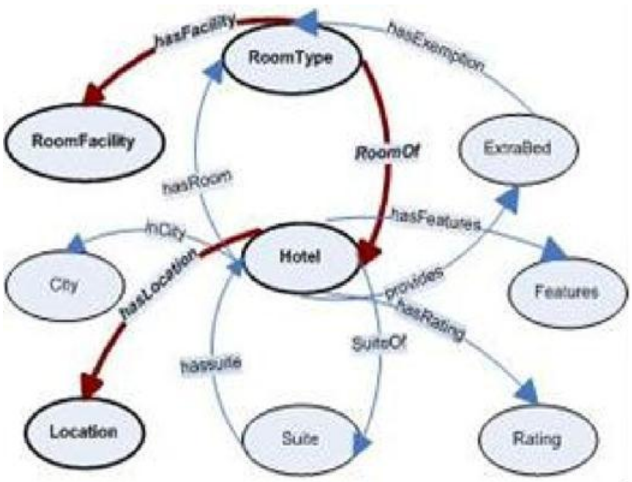

Fig. 4 Shortest path for the given query.

DOI:10.5963/IJCSAI0303005

Such queries are often posed due to confidence in the system and this confidence increases in querying for information using domain specific interfaces for search.

4) Pattern Identification

We have used the term pattern to represent the type of resources identified in the user query with positions for the exploration of filter value with the datatype property. Each filter expression is applied to the nearest datatype property.

5) DL Query Construction

An algorithm is used to construct triples from the QueryStructure. The DL generation algorithm uses the QueryStructure to form the target query by combining variables and triples. The DL Queries thus constructed serve as the engine for the extraction of information from the HTML pages.

VI. CONCLUSIONS

The current web provides data that is only loosely connected and can not be used for rich querying. The model presented in this paper demonstrates the effectiveness of the semantic web for searching and extracting information. The semantic web, enabling meaningful search, provides us with a tool to extract fewer and more relevant data from the vast World Wide Web.

REFERENCES

[1] S Ahmed, F.F, Hussain, S.F, Hameed, S., Ali, S.M.," Semantic Web E-Portal for Tourism". In IEEE DICTAP2012, Bangkok, Thailand, 2012

[2] WTO, World Tourism Organization, 2005.

[3] Cardoso, J. E-Tourism: Creating Dynamic Packages using Semantic Web Processes, W3C.

[4] Berger, H., Dittenbach, M. and Merkl, D, "Querying tourism information systems in natural language". Proceedings of the 2nd International Conference on Information System Technology and its Applications ISTA, Kharkiv, Ukraine. *200*3

[5] Blomqvist, E, 2009. *Semi-automatic Ontology Construction based on Patterns*. PhD thesis, Linköping University, Department of Computer and Information Science at the Institute of Technology.

[6] Choi, C, Cho, M., Choi J, Hwang M., Park J., Kim, P. 2009 *Travel Ontology for Intelligent Recommendation System*. Proceedings of the 2009 Third Asia International Conference on Modelling & Simulation, IEEE computer society, pp. 637-642.

[7] Embley,D., Tao, C, Liddle, C. 2002 *Automatically Extracting Ontologically Specified Data from HTML Tables of Unknown Structure.* Proceeding of the 21st Int. Conf. on Conceptual Modeling (ER02), Tampere, Finland.

[8] Hasany, N, Jantan, A.B, Selamat, M.H.B, and Saripan,. M.I, *Querying ontology using keywords and quantitative restriction phrases*. Inform. Technol. J., 9: 2010. pp. 67-78.

[9] Maedche, A, Staab, S. 2002 *Applying semantic web technologies for tourism information systems*. In K. Wöber, A. Frew, M. Hitz (Eds.). Proceedings of the 9th Int. Conf. for Information and Communication Technologies in Tourism, Springer Verlag.

[10] Niemann, M, Mochol, M, Tolksdorf. R, *Enhancing Hotel Search with Semantic Web Technologies.* JTAER 3(2):2008. pp 82-96.

[11] Niemann, M, Mochol, M, Tolksdorf. R, 2006 *Improving online hotel search - what do we need semantics for?* In Proc. of the Semantics

[12] Pan, B, Litvin, S.W, and Goldman, H.2006 *Real Users, Real Trips, and Real Queries: An Analysis of Destination Search on a Search Engine'*, Travel and Tourism Research Association Conference Proceedings. Dublin, Ireland

[13] Shi, F, Li, J, Hu, J. 2009. *Integration of Travel Information based on Information Extraction and Instance Matching. In:* Proceedings *of CSWS, Nanjing, China*

DOI:10.5963/IJCSAI0303005