Design and Implementation of an Aggregation-based Tourism Web Information System

Ainie Zeinaida Idris and Nor Adnan Yahaya

Malaysia University of Science Technology, Unit GL33 (Ground Floor), Block C,
Kelana Square, 17, Jalan SS 7/26, 47301 Petaling Jaya, Selangor Darul Ehsan, Malaysia

Summary
This paper discusses the design and implementation of a prototype web information system that uses web aggregation as the core engine. In this context, web aggregation is referred to as the process of integrating data extractable from various heterogenous web sources to meet certain new purposes. This prototype, referred to as MyTourism is developed by making use of Kapow Mashup Server which supports automated data extraction from web sources that can subsequently be stored either in the form of XML files or relational databases. MyTourism uses the tool to periodically extract data that are considered to be of interest to potential tourists, from selected Malaysian tourism websites. These extracted data are stored in Microsoft SQL Server 2005 for use by a simple aggregator that was developed using ASP.NET and C# in order to provide the intended data integration for the target users. This paper also presents our naïve technique to handle semantic conflicts among these extracted data and discusses the potential use of Semantic Web technologies for this purpose as part of our future works.

Key words:
Web data extraction, web aggregation, semantic web, Tourism Information System, Web Information System.

1. Introduction

Most of our travel and vacation information needs today can be satisfied by searching and browsing the Web. Hundreds of tourism web sites have been set-up by organizations, service providers or agencies in the tourism industry as their tourism information systems (TIS). We may end up spending the whole day surfing hundreds of web sites and get overloaded with un-necessary information. We then need to figure out how to process the information to suit our specific needs. Wouldn't it be much simpler to have only one system that could seamlessly interact directly with those web sites and provide all the information we need without browsing thru hundreds of web sites?

With the advent of powerful web data extraction tools, extracting and integrating data from these various web sites has now becoming easier and easier. This trend has given rise to a new phenomenon called web aggregation [1][2] which refers to the process of collecting information from a wide range of web sources. This has resulted in a great number of web aggregators, collecting information from various cooperating or non-cooperating sources.

The above arguments have motivated us to explore the problem of developing tourism web information systems that maximize the use of web aggregation to deliver the required data from existing tourism-related web sites to suit new purposes. This has resulted in a prototype known as MyTourism that we are going to describe in the subsequent sections. We have designed MyTourism as a combination of different tourism products, which include travel packages, accommodations, events, travel agents and destinations.

2. Overview of the Approach

Automatic processing of Web-based information is a complex task where unlike conventional data sources, data on the Web is in unstructured format, potentially unbound, and highly volatile [3]. To structure the process of building information aggregation application based on Web data for MyTourism, the overall tasks is broken down into the following sub-tasks:

- Task 1: Acquiring the required data from the source TISs,
- Task 2: Extracting the relevant data from various pages of the TISs,
- Task 3: Integrating, transforming the aggregated data, and
- Task 4: Delivering the data into "personalized content" to users via the MyTourism User-Interface.

Hence, central to the development of myTourism is the use of a reliable and easy-to-use web data extraction tool. We have chosen Kapow Mashup Server software [4] for this purpose where several Kapow software "robots" were developed to extract data from the identified Malaysian tourism related websites and store them in a single Microsoft SQL Server 2005 database.

A robot can be thought as a small program that has been programmed with the control flow and processing needed

Manuscript received December 5, 2009
Manuscript revised December 20, 2009

to support the intended data extraction from the target web pages as well as simple manipulations on them. It navigates through the target web pages of the corresponding TIS and extracts the required data into objects based on models created through Kapow's ModelMaker. These objects are then mapped into relational tables which allows for storing of the extracted data into conventional relational databases.

The other components of MyTourism were primarily developed using Microsoft Visual Studio .Net as the main development tool. Its application services include aggregating the extracted data and deliver personalized-content to travelers.

3. Architectural Design

MyTourism is basically structured based on the following Four-Layer Model (Fig. 1).

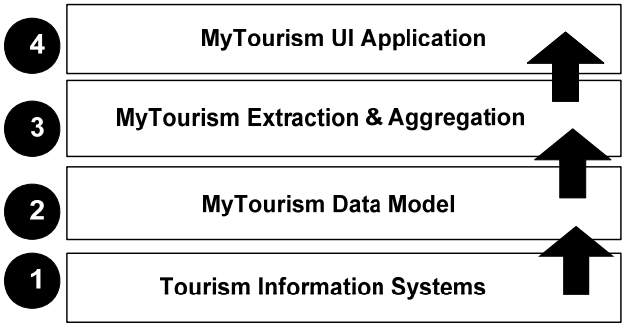

Fig. 1 MyTourism Four-Layer Model

.

Layer 1: Tourism Information Systems

Typically, a Tourism Information System (TIS) provides travelers with information such as accommodations, prices, the availability of travel packages and recommended destinations. The preliminary aggregation tasks are to understand the information architecture of each TIS and to identify relevant data to extract.

Layer 2: MyTourism Data Model

This is where the object models for use in web data extraction reside. The extracted data then is stored into a single database. Each object schema corresponds to a relational table in the database.

Layer 3: MyTourism Extraction and Aggregation

This layer comprises of Kapow robots and the Aggregator component. Our aggregator uses a simple *semantic mapping* mechanism applied for integrating those data objects, which involved the following steps:

1) Name the data objects that have the same kind of information but from different TIS with the same naming convention. For example, data objects that store a list of attractive vacation destinations in states are named as XX_Destination, where XX denotes the TIS abbreviation name.
2) Identify the relationship between those data objects and construct a single repository schema that integrates all those objects in a relational database, namely **MyTourismGSD**.
3) Build mapping rules:
   - To identify similar data from different sources. Based on the rules, only a single data from one of the sources is stored in MyTourismGSD.
   - To identify the data relationship with other data object(s).

Layer 4: MyTourism Application

This layer contains user interface component (i.e. web pages) and the Search Service component. The Search Service delivers a set of relevant information queried by a user, rather than have the user to search the information one after another. For example, a traveler who search for keyword "Diving in Malaysia", the search result shall not only return a list of destinations in containing the keywords "diving". In addition, it also returns list of travel packages, accommodations and other relevant information to each of the diving spots.

Searching for related information (or concepts) instead of or in addition to keywords shall improve efficiency and usability of this search service, i.e. through performing this type of semantic search. This service leverages on the semantic network of the data in MyTourismGSD database, which is defined by the Semantic Search Rules.

Fig. 2 shows the system architecture for MyTourism that supports the Four-Layer Model described earlier.

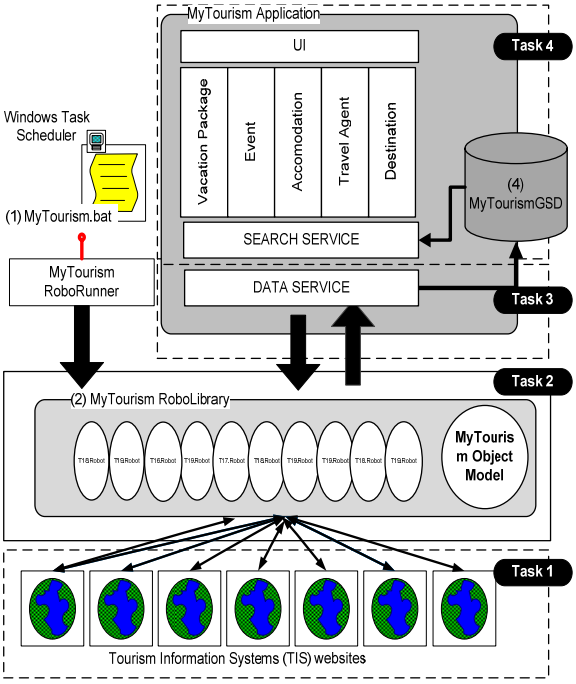

Fig. 2 MyTourism System Architecture

Using Kapow's RoboRunner, MyTourism robots are executed in batch in conjunction with Microsoft's "Scheduled Tasks" on daily basis through **MyTourism.bat (1)**. These robots and object models reside in **MyTourism RoboLibrary (2)**.

The next step is to integrate and relate the data from the heterogeneous TISs. In order to manage semantic heterogeneity in MyTourism database, the meaning of the interchanged information has to be understood across the entire solution. We use a set definition mappings or mapping rules to define or identify the correspondence relationship between the data. We have created a **Data Mapping Service (3)** for integrating the information in the objects into a consolidated repository, **MyTourismGSD (4)** database. In addition, a naïve Semantic Search technique is implemented thru **Search Service (5)**, to search for corresponding entries based on search term specified by the users. The information is then presented to the users on the user-interface **UI (6)**.

4. Component Model and Implementation

MyTourism comprises of four (4) sub-systems as depicted in Fig. 3 which are:

- Extractor
- Scheduler
- Aggregator
- MyTourism Application

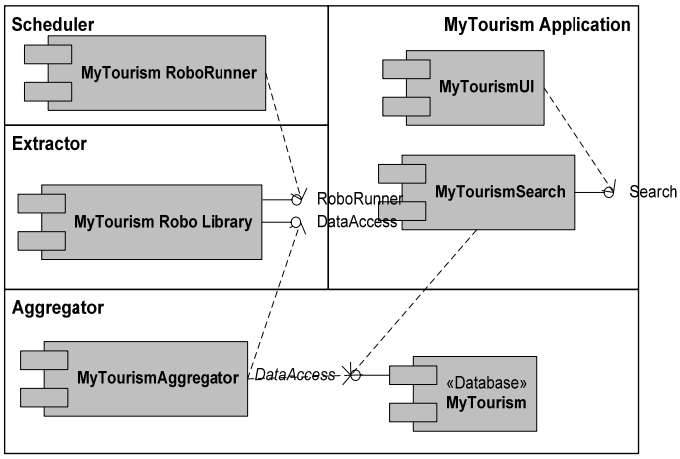

Fig. 3 MyTourism Component Diagram

4.1 Extractor

The Extractor subsystem utilizes a robot library, namely ***MyTourism*** library that serves as the deployment component for the robots. The data extracted by the robots are stored in the corresponding data records in MyTourism database. Essentially, MyTourism Robo Library contains :

- MyTourism.model which comprises of MyTourism object schemas. The data extracted by the robots are stored in the corresponding data records in MyTourism database.
- A collection of MyTourism robots.

4.2 Scheduler

The Scheduler subsystem comprises of a batch file (MyTourism.bat) which invokes Kapow RoboRunner commands to run the Kapow robots in batch on a daily basis. These robots are executed sequentially.

4.3 Aggregator

This Aggregator subsystem provides a service to extract and integrate data in MyTourism database into the consolidated repository, i.e MyTourismGSD database. In doing this, semantic heterogeneity issues are also managed. This aggregation is initiated by the Extractor once it has completed executing the robots. This subsystem comprises of two components : MyTourismAggregator and MyTourismGSD.

4.3.1 MyTourismAggregator

This component fetches data in the data objects resided in MyTourism database and aggregates the data into the corresponding data objects in the consolidated database, MyTourismGSD. It evaluates data values in the source

data object (i.e. MyTourism) against a set of pre-defined Mapping Rules to discover the mappings between the data sets. It then transforms the source data object's data value according to the set of pre-defined mapping rules and Synonyms. The Synonyms is a list of set of words which have same meaning or referring to the same meaning. This mapping approach leverages on regular expressions to annotate similar patterns between the two values (i.e. source data and target data), which is implemented through a class method. It utilizes the information in the Mapping Rules and the Synonyms to identify the corresponding output fields and the mapping of the output fields to sub-sections of the parsed value of the source data and transfer the matched data to the destination object. The extraction is done in a sequential manner as to determine the relationship between those data objects.

4.3.2 MyTourismGSD

MyTourismGSD is a database component on Microsoft SQL Server 2005. As a set of data is added into the corresponding data object in MyTourismGSD, a trigger is executed on the data object to update an index Mapping File named MyTourismKWord. It contains a list of titles and descriptions of data added into the corresponding data object in MyTourismGSD, which will be used by MyTourismSearch component.

4.4 MyTourism Application

MyTourism Application comprises of two main components, i.e. MyTourismSearch and MyTourismUI.

4.4.1 MyTourismSearch

MyTourismSearch Visual.Net C# class component provides the search functionalities in MyTourism, adopting a simple semantic search process as shown in Fig. 4. The user provides the Search with a phrase which is parsed from MyTourism search page. The Search uses regular expression analysis pre-configured with a set of rules that enables it to denote an object about which the user is trying to gather/research information and identify the objects or information which is semantically related to the denoted object. The rule set defines other relevant objects that the phrase is matched against title or description attributes of an object. The pre-configured rules defined are stored in an XML file, as shown in Fig. 5. For example, a keyword "redang package" is matched against destination name "Redang Laguna Beach Resort' in Destination data object, and consulting the Search Rules, the search result would also return the following information:

- List of accommodations available in Redang Island
- List of travel packages to Redang Island
- State where Redang Island is located

This component contains a method that crawl MyTourismKWord file and performs regular expression analysis to find a matched expression in the file against the parsed phrase. The result would be a data set of data objects returned back to the Search Result page. At the same time, the result is parsed to another method, to construct queries on the relevant object(s) detected for retrieving detail information on the corresponding search result.

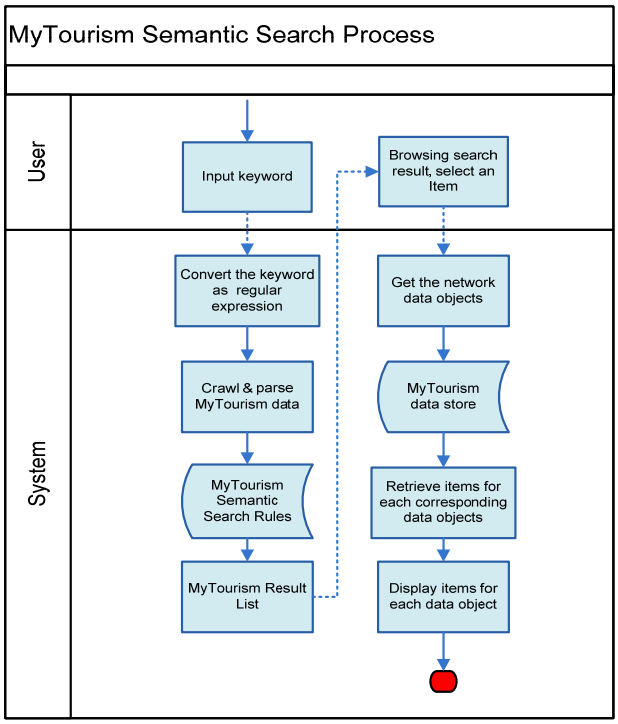

Fig. 4 MyTourism Semantic Search Process

```
<?xml version="1.0" encoding="utf-8" ?>
<SearchRules>
  <Rule>  <DataObject>State</DataObject>
<RelatedObject>Accomodation;Destination;Event;TravelPackage</Rela
tedObject>  </Rule>
  <Rule>  <DataObject>Destination</DataObject>
<RelatedObject>Accomodation;State;Event;TravelPackage;MalaysiaWe
ather</RelatedObject>  </Rule>
  <Rule> <DataObject>TravelPackage</DataObject>
<RelatedObject>State;Destination;Event;TravelPackage</RelatedObject
>  </Rule>
</SearchRules>
```

Fig. 5 Search Rules

4.4.2 MyTourismUI

MyTourismUI is the user-interface component that contains web pages for the users to interact with MyTourism application.

5. Challenges and Future Enhancements

As discussed earlier, one of the key challenges in the e-tourism is the growing need to aggregate information from the multiple TISs into one single portal for tourism and travel information. The implementation of myTourism shows that data extraction can be automated, converted into a structured form, stored in a consolidated database, and finally transformed into a format suitable for use by MyTourism application. However, this is a complex task as the extracted data are heterogeneous. Several problems were encountered in the aggregation process, as discussed in the following sections.

1. Different naming

This refers to the case when the same piece of information is being addressed with different names. For example, in the simplest case, a state is referred with different names by the source TISs such as the website http://www.tourism.gov.my refers Melaka state as 'Melaka', whereas the website http://www.malaysia.sawadee.com refers it as 'Malacca'. Generally, two schema elements in two data sources may also have the same intended meaning, but using different names. Thus, during integration, it should be realized that these two elements actually refer to the same concept.

In MyTourism this type of problem is handled by having a set of pre-defined Mapping Rules and a pre-defined Definition of Terms used in the TISs are applied. These two also form the basis for the integration process in MyTourism.

2 Structural Conflicts

There are often cases where the web page of interest contains data items with attributes that are although similar but not uniformly structured as shown in Fig. 6. In yet another case, same type of information may be provided on different levels of abstraction within their respective pages. For example, the list of golf courses in Kuala Lumpur page is located in different tag path from the list of golf courses in Kedah page. This scenario is very complicated. The robot created must be able to locate the two tag path of the "Golf Course" in the two respective pages. In addition, the list of golf courses in other pages for other states may also located in different tag paths. To create one robot that can cater for this scenario is very difficult. The work around is to create a robot that extract the information for each state, i.e. to create fourteen robots, and this sounds impractical.

Next, there are also cases where the data items are of the same type or referring to the same concept, but they are represented in different structures in their respective pages. For example, Travel Packages in www.tourism.gov.my is having different information structure with Tour Packages in http://www.virtualmalaysia.com. In a another case, the same information may be represented with different levels of granularity or details.

For example:
http://www.tourism.gov.my/en/destinations/item.asp?item=alor_star_tower
<Destination>
<Description>
<Key Tips>
<How to get there>
<Who to contact>
</Destination>

http://www.malaysia.sawadee.com/kualalumpur/places2.htm#petronas
<Places of interest>
<Description>
</ Places of interest >

Fig. 6 Attributes not Uniformly Structured

Issues related to *structural conflicts* can be resolved by creating more **robust** robots. **Robustness** is the term used to describe how well robots cope with web site changes. The more changes the robot can deal with (and still work correctly), the more robust it is [5]. However, creating a robust robot involves analyzing the web site in question, and to understand how it responds in various situations, such as the different abstraction level scenario illustrated above. As one of the tips given in (Kapow Technologies),

writing robust robots involves a kind of reverse engineering of the web site logic, and usually the only way to do this is through exploration.

2. Semantic conflicts

Some of the structural conflicts encountered are also due to semantic heterogeneities that are contextual and ontological in nature. In most cases, contextual heterogeneities are due to differences among local definitions, such as attribute types, formats, or precision. These differences can be easily resolved through Kapow RoboMaker by using its error handlers function and data converters.

On the other hand, ontological heterogeneities are mostly due to disparities in the interpretation or meaning of schema elements in different data sources. MyTourism current use of pre-defined mapping rules is not adequate to handle most of ontological heterogeneity cases. Hence, as part of future enhancements on MyTourism, the ontology based approach [6] to semantic heterogeneity resolution will be explored.

In addition, using ontologies as a tool for integrating TISs data also allows for semantic search methods based on the ontology. MyTourism Search Service can be enhanced by adopting the approach presented in [7], a methodology for implementing semantic search in an integrated biological data warehouse. Even though the data domain is different, but the fundamental ideas behind this approach can be adopted by MyTourism. Furthermore, the use of metadata will also help in improving the semantic search component of MyTourism.

Recently, ontologies have become the central focus of the Semantic Web initiative, giving rise to ontology description languages and associated technologies. MyTourism is already XML-based since it relies on Kapow as the main extraction tool where the latter produces extracted data in XML format. The incorporation of ontologies into MyTourism will allow us to capitalize on key XML-based technologies that are related to the future Web, i.e the Semantic Web. These include the use of the Resource Description Framework (RDF) [8] and Web Ontology Language (OWL) [9] for handling the metadata and ontology components of MyTourism, respectively.

6. Conclusion

In this paper, we have described the design and implementation of a prototype aggregation-based tourism web information system. The prototype called MyTourism, focused on Malaysian Tourism websites as the source for information extraction. The implementation of myTourism shows that data extraction can be automated, converted into a structured form, stored in a consolidated database, and finally transformed into a format suitable for use by tourism related applications. Essentially, this will help in minimizing efforts in developing contents from scratch. However, the inherent heterogeneity of data in the source websites has made the integration and repurposing of the extracted data less straightforward. We have presented some of these challenges and highlighted the potential use of Semantic Web technologies in improving this prototype.

References

[1] *Aggregator*. (2009, March 27). Retrieved April 27, 2009, from Wikipedia, The free encyclopedia: http://en.wikipedia.org/wiki/Aggregator

[2] Madnick, S. E. and Siegel, M (March 2002). Seizing the Opportunity Exploiting Web Aggregation. *MISQ Executive, Vol. 1, No. 1, pp. 35-46.*

[3] Abitebou, S., Buneman, P., & Suciu, D. (1999). *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann

[4] http://kdc.kapowtech.com/documentation_6_5/Technical/TechnicalDataSheet6_5.pdf

[5] Kapow Technologies . In *RoboMaker User's Guide Kapow Mashup Server 6.4* (p. 144). Kapow Technologies.

[6] Hakimour, F., & Geppert, A. (n.d.). Resolving Semantic Heterogeneity in Schema Integration: An Ontology Based Approach

[7] Cao, S.-L., Qin, L., He, W.-Z., Zhong, Y., Zhu, Y.-Y., & Li, Y.-X. (2004). Semantic Search Among Heterogeneous Biological Databases Based on Gene Ontology. *Acta Biochimica et Biophysica Sinica 2004, 36(5)* , 365-370.

[8] Manola, F. and Miller, E. (Eds), RDF Primer, W3C Recommendation, 10 February 2004. Available at : http://www.w3.org/TR/rdf-primer/

[9] McGuiness, D. L. and Harmelen, F. V. (Eds), OWL Web Ontology Language Overview, W3C Recommendation, 10 February 2004. Available at : http://www.w3.org/TR/owl-features/

Ainie Zeinaida Idris obtained her Bachelor of Science (Computer Science) from Universiti Sains Malaysia and currently completing her Master of Science in Information Technology at Malaysia University of Science and Technology (MUST).

Nor Adnan Yahaya obtained his PhD in Computer Science from Northwestern University, USA in 1987. He is a Professor of IT at Malaysia University of Science and Technology (MUST). His current research activities are focused on the development of tools and innovative applications related to emerging Web technologies such as web aggregation, web services, web agents, and the Semantic Web.