International Journal of Advances in Computer Science and its Applications – IJCSIA
Volume 4 : Issue 1 *[ISSN 2250 – 3765]*
Publication Date : 09 January 2014

Uncertainty Issues in Automated Web Service Selection

Shailesh Pancham Khapre[*, 1], Dr. M.S. Saleem Basha [1], G. Sambasivam[1]andB. Saravana Balaji[2]
[1]Department of Computer Science& Engineering
Pondicherry University
Pondicherry, India.
[2]Assistant Professor, Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College, Coimbatore.
{shaileshkhaprerkl, gsambu, saravanabalaji.b, m.s.saleembasha}@gmail.com

***Abstract*—Present world is an era of automation, machines have been granted the power of decision making, so is the case in Information Science. Recent developments in web services have shown the capability of semantic web in decision making through automated Discovery & Selection. Though automated web service discovery is a huge achievement in information Science but there lies the uncertainty factor which hampers the accuracy of decision making thus making it difficult to deliver exact service. In this paper we proposed an experimental system to identify uncertainty issues related to automated Web Service processing and shown how these uncertainties can be reduced through Web Content Mining and User Preference Mining. The paper is concluded with the discussion of possible future enhancement in Web modeling standards for handling uncertainty Issues.**

***Keywords*—Web service, Uncertainity Issues, Information retrieval, User Preference.**

I. INTRODUCTION

Boom in Information Technology has led to tremendous use of web based services which proportionately lead to exponential growth of information or data over web. Accessing through this humongous data is a big challenge for web Search systems. Using this data to gather information and knowledge hidden in them can be advantageous for both industries & individuals. Hence there is an evolution of web search system to different private decision support systems ranging from marketing systems, competitors or price tracking systems.

A person can search for web service as per his demand, but this is a time consuming or tedious job, that's where Semantic Web[1] comes in to play. Semantic Web [1] main vision is to automate web search activities & processing. Using Semantic web will speed up the searching process, an can find wider range of resources and when needed soften or optimize search criteria. As per Uncertainty Reasoning for the World Wide Web (URW3) Incubator Group [2]: "…as work with semantics and services grows more ambitious, there is increasing obligation of the need for principled approaches for representing and reasoning under uncertainty. The word "uncertainty" is intended to incorporate a variety of forms of incomplete facts, including incompleteness, vagueness, inconclusiveness, ambiguity, and others. The phrase "uncertainty reasoning" is intended to represent the all available methods aimed towards representing and reasoning with knowledge where boolean truth values are unknowable, unknown, inapplicable or inappropriate. Common approaches for uncertainty reasoning include fuzzy logic, Dempster-Shafer theory, probability theory and several other methodologies." Our goal in this paper is to concentrate on the issues associated with transforming or minimizing human abilities or web interactions by software. With this view, some uncertainty arises "human_to_machine(mediator)_to_web" specific, like faulty sensors, input errors, medical diagnosis, weather prediction, gambling etc. These uncertainties are difficult to deal with for human alone and also outside the web.

According to Turtle and Croft [3], uncertainty in information retrieval originates especially in three areas: "Firstly, the problem of the representation and annotation of a resource (service). Difficulties also arise in case when attempting to represent the belief degree to which a resource is relevant to the task. The second problem is the way of representation of information, action; which a user needs to retrieve or process. Thirdly, "association of user needs to resource concepts".

As per our understanding, these uncertainties also apply to our case, when transforming or minimizing human abilities or web interactions by software. A generic Schema for automated web service Discovery & tasks associated to these three problems are illustrated in Figure 1.

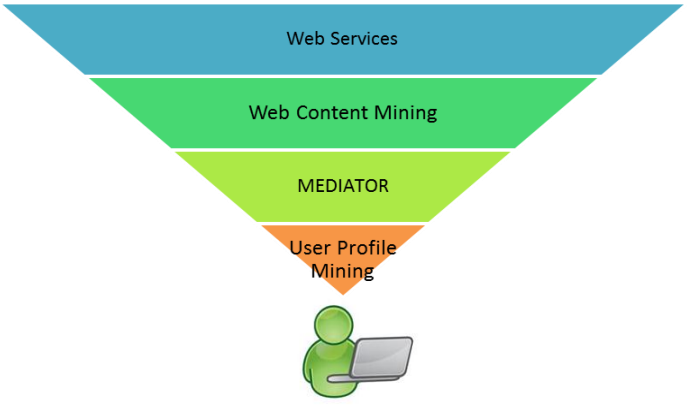

Figure 1. Schema for automated web service Discovery

Our goal is to discuss uncertainty issues based on a system integrating the whole Web Discovery & Selection process i.e. dealing with web & user. The uncertainty problem here is associated with two deductive procedures (Web & User). Two types of data mining appear in Automated Web Service Selection One is Web content mining and second is user profile or preference mining. A Mediator or Reasoning engine will perform the matching part and query evaluation and optimization.

A. A Driving example

As a driving example, assume that users are looking for a tour packages in a certain region. The amount of information is huge and distributed over several sites. Moreover users have their own preferences which are soft and problematic to express in a standard query language. From the Mediator viewpoint, there is no chance to evaluate user's query over all information. Let consider Mediator using Skyline threshold algorithm [7], which can find best answers that are not dominated by any other objects. Skyline algorithm works under following suppositions; First, there should be access to objects i.e. in our case tour packages grouped in different lists ordered by user particular attribute, which is bounded by a numerical score i.e. from 0 to 1, e.g. $f_1(x) = cheap(x), f_2(x) = days(x)$, ... Second, there should be a combination function computing total fuzzy preference value of an object based on preference values of attributes, e.g. $@(x) = \left(\frac{3*cheap(x)+close(x)}{4}\right)$.

In the real application we have to consider different users with possible different attribute orderings f_1^u, f_2^u and combination functions $@^u$. These exemplify the overall user preference $@^u(f_1^u, f_2^u)$ and the user profile for this task. The task of user profile mining part is to find these particular attribute orderings and the combination function (using user's ranking of a sample of tour packages).

On the server side, the information of vendors, companies or advertisement is very often presented in a structured layout containing data records. These structured data objects acts as a very important type of information for decision making systems dealing with competitor tracking, market intelligence or tracking of pricing information from sources like vendors.

This structured data is extracted and feed to our Mediator. Due to the huge size of Web, there exist a bottleneck of the degree of automation of data should be extracted. A balance should be maintained between the degree of automation of Web data extraction and the amount of user (administrator) effort which is needed to train data extractor for a special type of objects (increasing precision).

First restriction we make is that we consider Web pages incorporating several structured data records. This is basically the case of Web pages of companies and vendors containing information about products and services and, in our case, tour packages. Main difficulty arises in extracting data and especially attributes values to Mediator.

Our main contributions are:

- Identification of uncertainty issues in web content mining system and while extracting attribute values from structured pages with several records
- Identification of uncertainty issues in user profile model
- Discussion of coupling of these systems via a Mediator based on Skyline threshold algorithm complemented by various storage and queryingmethods.

II. UNCERTAINTY IN WEB MINING

In this section we describe our understanding with a mining system for information extraction from structured web pages and try to point out places where uncertainty rises.

Using our driving example, imagine a scenario where user looking for a tour packages in a certain location. A relevant page for a user searching for tour packages can look as on figure 2. As per the user point of view comparing more similar pages would increase the chance of finding the best tour package. But this is both exhausting and time taking job. An automated tool would enhance this search.

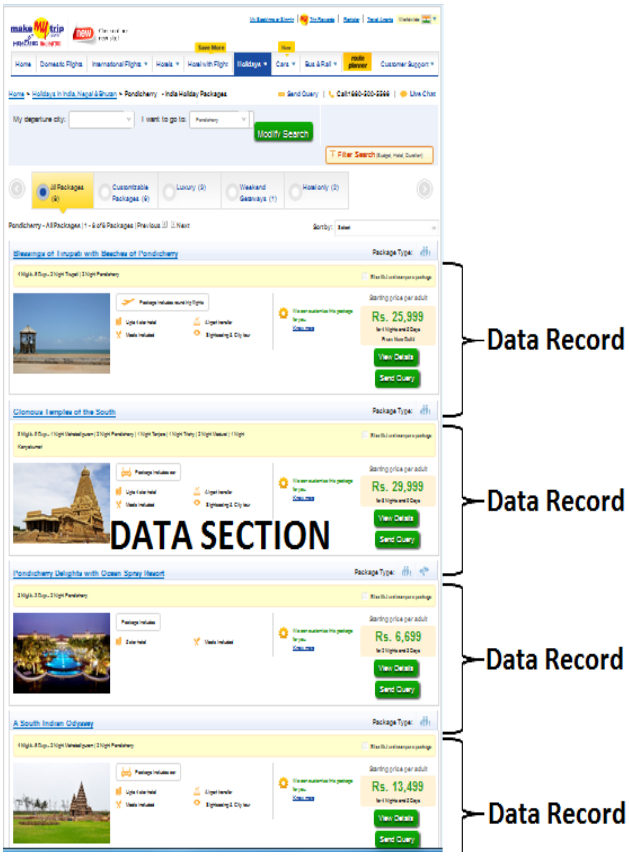

Figure 2. A Generic Web Page Containing Records

Semi- automatic extraction systems like Lixto [5], Stalker [9] or WIEN [8] can be used for structured Web data extraction. The primaries required are user pre-annotated pages, which are used for training process. Moreover, they are most appropriate for pages, which have dynamic content with fixed structure.

Our clarification is based on different approach. Instead of training techniques we use automatic discovery of data regions which will encompass multiple similar data records on the page which is supported by an extraction ontology [9], it is used to extract the values from data records. There are many ways to search for similar records in source tree. The system IEPAD [10] uses the Patricia tree (radix-tree) to catch the repeating sequences. The MDR system [11] operates directly on the DOM tree of input in which it searches for repeating node sequences with same parent. However, both methods have same goal, search objects of interest in the whole web document. Which is time consuming and, what we have experienced, it surprisingly decreases precision. Furthermore, these systems do not extract attribute values from data records.

In this paper we describe a system as a sequence of both data record extraction and attribute value selection, with likelihood of ontology starting almost from scratch (e.g. user search key words).

The proposed system will be described in several phases, which are described in the following sections.

A. *Discovering Data Regions and Data Records*

The first step in the extraction process is to retrieve all the relevant web pages. Egothor .v3 [12] can be used for automatic localization of such resources, it is an open-source, high-performance, full-featured text search engine. This system is used for downloading the HTML source codes of relevant pages.

The next step is to build a DOM model of the concern web page. DOM model is used for extracting both data region and data records. Figure 2 shows an example of relevant web page. This page contains summary information about three Tour packages, i. e. three data records. All of them incorporated in a single data region. Our main goal is to automatically discover this data region and records within. One should note that the discovery process is not limited to the single-region pages.

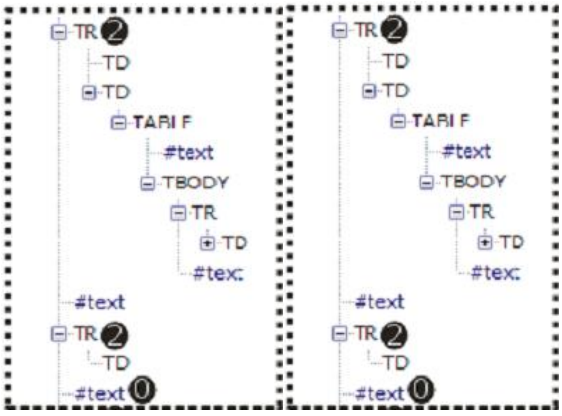

Figure 3. Dom Subtree of Records

Input DOM tree can be prune by omitting elements which do not carry any textual information in their subtrees this will reduce the search space and will increase precision.

An example of such tree is shown on the Figure 3 – the black circles carry the number which represents the relevance of the particular node. Nodes having zero relevance factors are omitted from the data record search. Here arises the first Uncertainty issue i.e. to identify nodes with relevant information in the sub-tree.

Next, we use breadth first tree alignment to detect data regions and records by taking element tuples, triples etc. and comparing their corresponding subtrees by Levenshtein distance metrics. Uncertainty arises when measuring the relevance of similar tags i.e. to adjust the similarity measures for discovery of similar tags.

Often every repeated sequence of tags learned makes up a real data record (a single tour package). All attributes of this record can be found in one sub-tree which can be feed to the extraction ontology to retrieve attributes. However, the detached data records can pose a problem in the region discovery phase.

Typically a data record constitutes a single visual region, which means that attributes of these records have a common subtree. It is therefore necessary to identify detached data records and separate attributes of these records which is an uncertainty problem.

B. *Attribute Values Mining*

Ontology is used to extract the actual attribute values of product in the web page. This ontology is dynamic – as it starts from the scratch, enclosing user search keywords, and subsequently it evolves with new keywords and values (using standard vocabularies). OWL syntax is used to incorporate additional annotation properties and allows the specification of values extraction parameters: e. g. a use of regular expression to match the attribute values, an unambiguous enumeration of possible attribute values, or the tuning parameters such as maximum or minimum attribute value length.

An example of ontology specification can be seen on Figure.4

```
<owl:DatatypePropertyrdf:ID="hasPrice">
<rdfs:domainrdf:resource="#Package"/>
<p1:maxLengthrdf:datatype="http://www.w3.org/2001/XMLSchema#strin
g">    10  </p1:maxLength>
<p1:patternrdf:datatype="http://www.w3.org/2001/XMLSchema#string">
(\$)? ?[\d]{1,10} ?(.){1,3}  </p1:pattern>
<rdfs:labelrdf:datatype="http://www.w3.org/2001/XMLSchema#string">
PRICE   </rdfs:label>
</owl:DatatypeProperty>
```

Figure 4. An example of ontology

The extraction process can be enhanced in various ways. A richer Ontology can help retrieve more relevant output. Additionally the approximate regular expression matching algorithms can be employed, which allows to identify and repair mistyped or mismatched attribute values.

III. MEDIATOR

A. Semantic Web Architecture

Mining user preference is done by the service provider locally and it is assumed that mined data are stored in Mediator. Mined data have to be modeled on an open world assumption model (OWA), in this case using traditional database models are inappropriate. That's why RDF databases are used to preserve the relations & to represent attributes & values related to object. A typical schema of record resembles a RDF statement.

Resource	Attribute	Values	Extracted_from	Extracted_by	Using_Ontology
Package 1	Price	V1	URL1.html	Method_1	O1
Package 1	Days	D1	URL1.html	Method_1	O1

We have not attached records with any uncertainty degree. We can evaluate it according to the remaining values (e. g. it can be known that Method1 is highly reliable on extracting price, but less on days). Gathering or collecting user interest is important to find out what we are looking for and which attributes values to be extracted. For Mediator we need to know the ordering of particular attributes and the combination function.

B. User profiles as the user preference model

Mining user preferences is a good idea to deliver Web Services satisfying user need. One way to model user preferences is to use user profiles. Consider we have a set of user profiles $P_1, \dots P_k$ and we know the ideal Tour package for each profile. These profiles can be created as the clusters of users or manually by an expert in the field. Independent of the way profiles are created, we have ratings of Tour packages associated with each profile, thus knowing the best and worst packages for that profile.

So the days d_i of user User1 profile U1 from each profile U_1 is computed as

$$d_i = \frac{\sum_{j=1,\dots,n} \left|\text{Rating}(User\ 1, o_j) - Rating(P_1, o_j)\right|}{n}$$

Equation (1) represents the average difference between the user's rating of an object o_j and profile's P_i's rating.

The ideal tour package for the user can be computed as an average of ideal tour packages for each profile P_i, weighted by the inverse of distance di. The average is computed on attributes of tour packages. Formally,

$$IdealTour(User\ 1) = \frac{\sum_{j=1,\dots,n} IdealTour(P_i)/d_i}{\sum_{j=1,\dots,n} 1/d_i}$$

Then, *IdealPackage*(User1) is the weighted centroid of profiles' best Tour packages. An example of data, user profiles' best tour package and user's best tour package is on Figure 5. User's best Tour package is clearly closest to Profile 3.

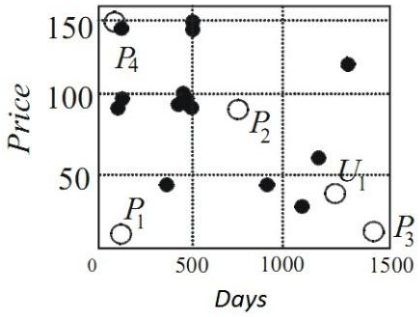

Figure 5. Positions of best tour packages for user profiles and for user

After the *computation* of the ideal Package tour for the user, we will use the output for computing ratings of remaining Tour Packages. Advantage of this user model is that it can be used in the Skyline threshold algorithm.

IV. UNCERTAINTY IN USER PREFERENCE EXTRACTION

As per our understanding, user preferences are expressed in the form of taxonomical rules, where the values of attributes are assigned their grades as per the orderings of the domains of these attributes. The higher the grade, the more suitable (preferable) the value of an attribute is for the given user. This form of grading links to truth values well-known in fuzzy reasoning and thus the orderings resembles fuzzy functions.

Fuzzy logic can be used to formulate combination function using fuzzy aggregation function [14]. Main assumption of our knowledge of the user preferences is that we have a (fairly small) sample of objects (Tours Packages) assessed by the user. Goal is to gather user's preferences from this sample estimation. The idea is to use this learned user preference to extract top-k objects from a much larger amount of data. Moreover, using the user sample estimation, we do not have to face the problem of matching the query language and document language.

There are many approaches for user preference modeling; the most used is collaborative filtering technique [13]. Our technique is content based filtering – it uses information about attributes of objects.

A. Local Preferences Learning

There are several techniques of user's preferences learning of particular attributes (UNC5) represented by fuzzy functions $f_1, f_2, \dots$, on attribute domains. Many of them use regression methods. A problem with this technique is that there can be potentially a big number of Tour packages of one sort (e.g. Distance ones) but the discovery of user preference (Short, Medium or Long) should not be influenced by the number of such Tour packages. Regression normally counts number of objects.

Another approach without using regression is the following. The view of the entire domain of attribute Distance is in Figure 6. Our observation shows that with increasing price, the rating is decreasing. Using these methods we can extract local preference which can be described as a fuzzy function (here small, Short,…) and hence are usable for Skyline Threshold algorithm.

s

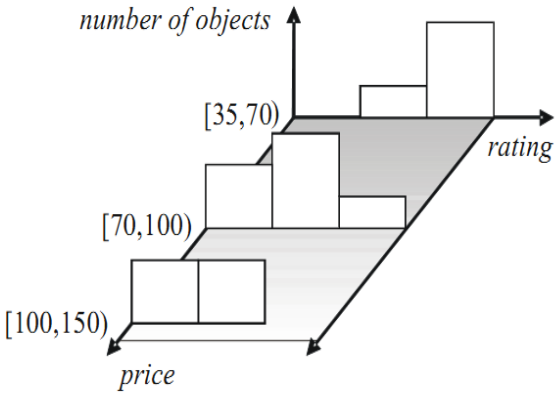

Figure 6. Overall Attribute Ratings

B. Combination Function learning

Skyline Threshold algorithm requires a combination function @, which integrates the particular attribute preference degrees $f_1, f_2, \ldots$, (local preferences) to an overall score – $@(f_1, f_2, \ldots,)$ – using which the top-k answers will be computed.

There are several ways to derive Combination Function, uncertainty commonly arise while learning phase which is because of the above discussed uncertainties. Combination Function derived is an instance of classification trees with monotonicity constraints. Inductive Generalized Annotated Programming (IGAP) can be used for aggregation function. The result is a set of Generalized Annotated Program rules in which the combination function has a form of a function annotating the head of the rule – here the quality of Tour Package:

```
User1_TourPackage(T) good in degree at least @( f1(x), f1(y), ...)
IF User1_TourPackage_price(x) good in degree at least f1(x) AND
User1_TourPackage_days(y) good in degree at least f2(y).
```

V. The Implementation

We proposed and implemented the Mediator system for performing top-k queries over RDF data. The system gathers information from local or Web data sources and combines them into one ordered list. Each time a user comes with different ordering so to avoid reordering, we have designed a general method using B^+trees for fuzzy ordering of a domain [15]. We have implemented classes for standard user scoring functions, and Skyline algorithm. Detailed description of implementation is out of the scope of this paper.

VI. Conclusions

Using an tentative implementation, in this paper we have identified several uncertainties arising, when classifying HTML nodes with relevant information in the sub-tree, tweaking similarity measures for discovery of similar tag subtrees, classifying single data records in non-contiguous html source, mining attribute values, Mining user's preferences of particular attributes, study the user preference combination function.

We have carry out trial with some solutions. One way is to use a Fuzzy description Logics (FDL) with both concepts and roles fuzzified. One problem of inserting FDL with fuzzy roles into OWL is that they consist of subject, predicate, object and the fuzzy value which cannot be directly modeled by RDF data. Second possibility is to use a FDL where only concepts are fuzzified and roles remain crisp (and hence both roles and fuzzy concepts can be modeled by RDF data).

VII. Acknowledgement

This work is a part of the Research Project sponsored under the Major Project Scheme, UGC, India, Reference No: F. No. 41-619/2012(SR), dated 1st July 2012. The authors would like to express their thanks for the financial support offered by the Sponsored Agency.

References

[1] T. Berners, J. Hendler and O. Lassila, "The Semantic Web" Scientific American Magzine, 2001.

[2] K. J. Laskey, K. B. Laskey and P. Costa, "Uncertainty Reasoning for the World Wide Web", W3C Incubator Group Report 31 March 2008.

[3] H. R. Turtle, W. B. Croft,"Uncertainty in Information Retrieval Systems" Uncertainty Management in Information Systems, pp:189-224, 1996.

[4] D.H. Kraft,G. Pasi and G. Bordogna "Vagueness and uncertainty in information retrieval: how can fuzzy sets help?" Proceedings of the 2006 international workshop on Research issues in digital libraries. vol. 3, 2007.

[5] R. Baumgartner, S. Flesca, G. Gottlob, "Visual Web Information Extraction," VLDB Conference, 2001.

[6] A. Giurca, "Using Uncertainty in Information Retrieval" http://*www*.informatik.tu-cottbus.de/~agiurca/papers/anale02.pdf

[7] J. Xin "Efficient threshold skyline query processing in uncertain databases" Natural Computation (ICNC), pp: 311 – 315, 2011.

[8] N.Kushmerick, "Wrapper induction: efficiency and expressiveness" Artificial Intelligence, vol. 118, pp.15-68, 2000.

[9] I. Muslea,S. Minton andC. Knoblock "A hierarchical approach to wrapper induction" Conf.on Autonomous Agents, 1999.

[10] C. H.Chang, S. L. Lui, "IEPAD: Information extraction based on pattern discovery" WWW-10, 2001.

[11] B. Liu, R. Grossman, Y. Zhai, "Mining Data Records in Web Pages." In: Proc SIGKDD.03, Washington, DC, USA, 2003.

[12] Egothor, http://www.egothor.org/

[13] R. Fagin, A. Lotem and M. Naor, "Optimal Aggregation Algorithms for Middleware." In Proc. 20th ACM Symposium on Principles of Database Systems, pp-102-113, 2001.

[14] C. Aggarwal, "Collaborative Crawling: Mining User Experiences for Topical Resource Discovery" , IBM Research Report, 2002.

[15] A. Eckhardt, T. Horváth, P. Vojtáš, "PHASES: A User Profile Learning Approach for Web Search." WI'07 Web Intelligence Conference, CA, USA, 2007.