Extracting Information from Semi-Structured Web Pages by Considering User's Context.

Conference Paper · January 2010

Source: DBLP

CITATIONS
0

READS
93

5 authors, including:

Mahmood Shakir Hammoodi
University of Babylon
15 PUBLICATIONS **51** CITATIONS
SEE PROFILE

Hamidah Ibrahim
Universiti Putra Malaysia
230 PUBLICATIONS **1,381** CITATIONS
SEE PROFILE

Ali A. Alwan
Ramapo College
126 PUBLICATIONS **730** CITATIONS
SEE PROFILE

Aida Mustapha
Universiti Tun Hussein Onn Malaysia
383 PUBLICATIONS **3,678** CITATIONS
SEE PROFILE

Extracting Information from Semi-Structured Web Pages by Considering User's Context

Mahmoud Shaker [1], Hamidah Ibrahim [2], Ali Alwan [3], Aida Mustapha [4], Lili Nurliyana Abdullah [5]
Department of Computer Science
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia, 43400 Serdang, Malaysia
[1] Mah222254@yahoo.com, [2] hamidah@fsktm.upm.edu.my, [3] ali83_upm@yahoo.com,
[4] aida@fsktm.upm.edu.my, [5] liyana@fsktm.upm.edu.my

Abstract - *Nowadays, many users use web search engines to find and gather information. User faces an increasing amount of various semi-structured information sources. The issue of correlating, integrating and presenting related information to users becomes important. When a user uses a search engine such as Yahoo and Google to seek a specific information, the results are not only information about the availability of the desired information, but also information about other pages on which the desired information is mentioned. The number of selected pages is enormous. Therefore, the performance capabilities, the overlap among results for the same queries and limitations of web search engines are an important and large area of research. Extracting information from the web data sources also becomes very important because the massive and increasing amount of diverse semi-structured information sources in the Internet that are available to users, and the variety of web pages making the process of information extraction from web a challenging problem. It is more challenging when an extracted information which is relevant to a user might not be relevant to other users. Thus, an information extraction that considers user's context more specifically user preferences would provide better results to the user. Thus, this paper proposed a framework for extracting information from semi-structured web pages by considering user's context.*

Keywords: Extracting Information, Web Pages, User's Context

1 Introduction

At the present time, the Internet is general and many people use the Internet to find information. A variety of web pages and the frequently changing of information in web data sources make searching and extracting information very difficult. When Internet users want to get information about hotel for example, they first visit search engines such as Yahoo and Google, and then visit all web sites suggested by the search engine.

Many queries processed on the World Wide Web do not return the desired results because they fail to take into account the context of the query and information about user's situation and preferences [11]. The use of context is significant in the interactive applications thereby limit the query results to only the most interest and the most relevant to the user's demand. It is particularly important for applications where the user's context is changing rapidly, such as in both handled and ubiquitous computing [2]. There have been a variety of definitions to define the context [2] [3] [4]. We choose the most appropriate definition given by Anind [2] where context is defined as "any information that can be used to characterize the situation of any entity (i.e., attribute). An entitiy is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves". In addition, a system is context-aware if it involves the context to return more accurate and relevant information and/or services to the user, where relevancy depends on the user's task. The problem associated with information extraction, in essence, arises because context is missing from the specification of the query.

In [1], we proposed an approach for extracting information from semi-structured web pages that handles genuine web tables, non genuine web tables, and synonym. The work presents in this paper is an extension of our work in [1] where user's context is being considered to be one of the factors that influences the results of the query. In this paper, we highlight the components of the proposed framework which encompasses User Preferences (UP) which is responsible to save the user's profile and preferences; Query Interface (QI) which is utilized by the query issuer to submit his request; Information Extraction (IE) this component is responsible to extract and classify the desired web pages that are obtained from QI and convert the extracted and classified web pages into text form; Relevant Information Analyzer (RIA) which is used for determining the relevant information extracted from Information Extraction (IE); and Match Relevant Information (MRI), in this component we match the relevant extracted information that have been collected with the user's profile and preferences to get the relevant results that meet the user's need.

The rest of the paper is structured as follows. In Section 2, the previous works related to this research are highlighted. Section 3 presents the proposed framework with its components. Finally, Section 4 concludes the paper.

2 Related Works

Many researchers have proposed different approaches for extracting information from semi-structured web pages as discussed below.

Srinivas, Fatih, and Hasan (2007) [7] work on information extraction from web pages using presentation regularities and domain knowledge. They argued that there is a need to divide a web page into information blocks, or several segments before organizing the content into hierarchical groups and during this process (partition a web page) some of the attribute labels of values may be missing.

There are many researches focusing on information extraction for business intelligence to collect available information for companies and easement the efforts concerned in gathering, merging, and analyzing information. Horacio, Adam, Diana, and Kalina (2007) [8] proposed the MUSING project (Multi-industry, Semantic-based next generation business intelligence), that needs to cover many semantic categories including locations, organizations and specific business events to help companies that want to take their business overseas and concerned in knowing the best place to exploit.

Fei and Zhuang (2005) [10] proposed an information extraction system that aims to automate the tedious process of extracting large collections of facts from large-scale, domain-independent, and scalable manner. The biggest of these challenges stems from the fact that search engines only make a small fraction of their results accessible to users.

Another stream of researchers work on extraction of information with agent. Sung, Kyung, Tae, and Hyuk proposed an Intelligent Traveler Support System (ITSS) (2001) [6] for helping traveler to find important information about traveling more easily and effectively. There are many limitations in the traveler supporting system. One of these limitations is the system deals with limited web pages which are related with destinations and weather. Thus, travelers need to search through the numerous web pages to gather all the necessary information by using search engines such as Yahoo and Google.

Gilles (2006) [9] proposed a new method for extracting information from the web by using wrappers and this method (wrapper construction) is based on different techniques: labeled page based induction, relation extraction, structure discovery and most recently a new approach based on the generalization of the contexts of a small set. The limitation is to build a wrapper for a data source which can extract a relation it contains. The description of the relation to extract is given in the form of a set of example instances.

However, from one side most of the previous approaches have concentrated on extracting information from semi-structured web pages without considering the user's context as one of the factors that might influence the results produced by information extraction. From the other side, there have been many approaches that have considered the user-context such as user profile, user preferences, user activities, etc. These work focused exclusively on the context-aware query processing ([3], [4], [5], [12], [13], [14], [15], [16]). Thus, an approach is needed for extracting information from semi-structured web pages that considers user's context as a filter to retrieve information that suits the user's interest.

3 The Proposed Framework

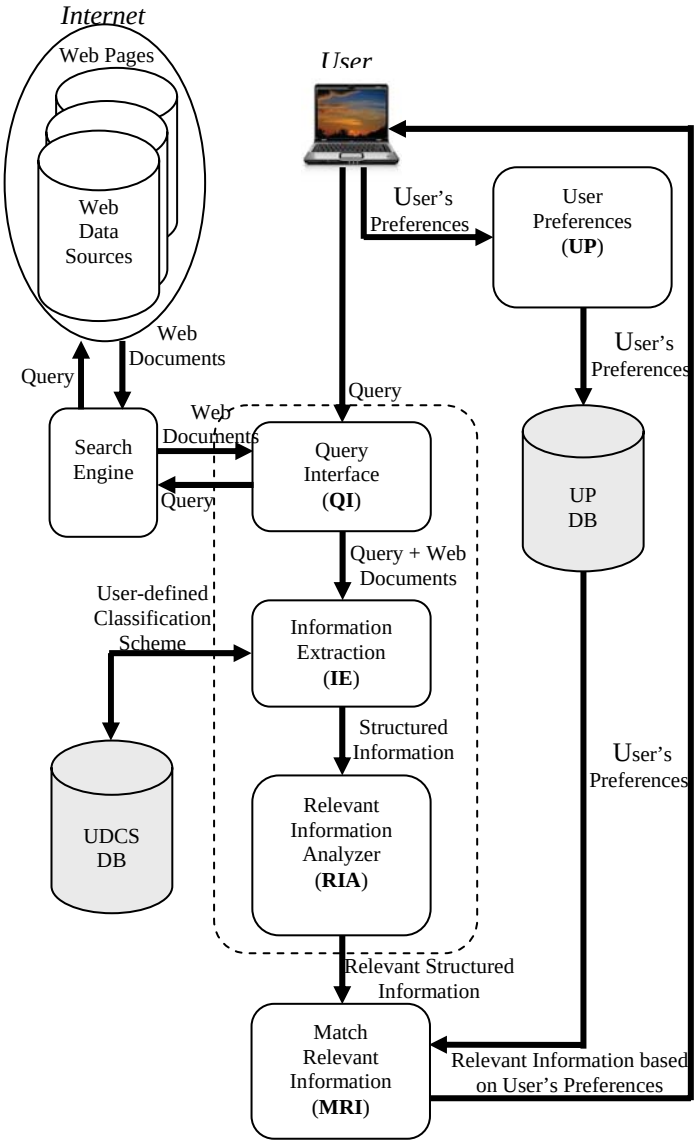

Figure 1. The proposed framework

Figure 1 illustrates the proposed framework for extracting information from semi-structured web pages with user's context. The framework consists of the following main components: User Preferences (UP), Query Interface (QI), Information Extraction (IE), Relevant

Information Analyzer (RIA), and Match Relevant Information (MRI). A user can define his/her own classification scheme consisting of attributes, sub-attributes, group of the sub-attributes that are relevant to the domain of interest. This information is stored in the UDCS database. For example, the attributes that might be stored for a hotel domain are: General, Location, Category, Room, and Facilities. In the following we discuss each of these components in more details. The discussion is based on the hotel domain which we have chosen as our case study.

3.1 The User Preferences (UP)

The UP component is defined to manage the user's profile and preferences. User is required to fill up his profile as well as his preferences. Example of a user's profile and preferences is shown in figure 2.

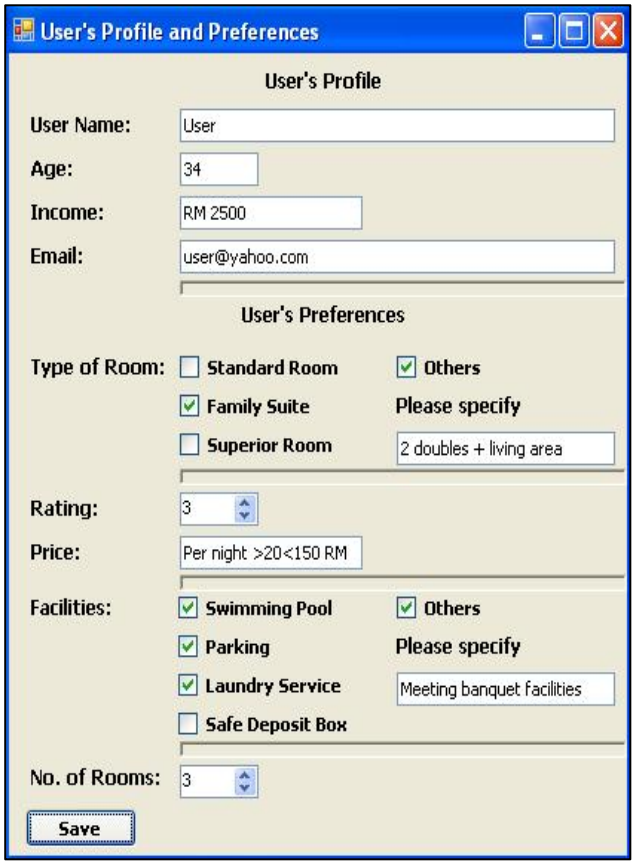

Figure 2. The user's profile and preferences interface

As our case study is related to the hotel domain, thus the preferences that can be gathered from a user include type of room, rate, price, facilities, etc. The user's profile and preferences are then stored in the UPDB to be accessed later by the MRI component.

3.2 The Query Interface (QI)

The QI is the key entry to the web and tool for accessing information. A user writes a query in the query interface, and the query is send to a search engine to search the web data sources. The web documents (i.e., the results of user's query received from a search engine) are sent to Information Extraction (IE) as HTM files. Example of a simple query is shown in figure 3.

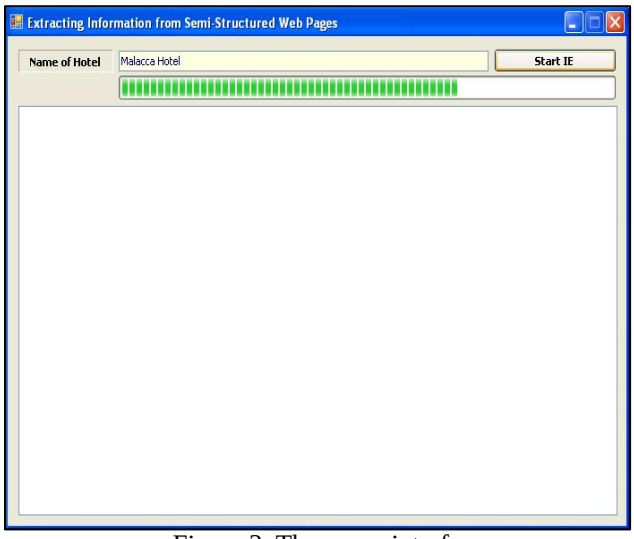

Figure 3. The query interface

Figure 4 illustrates examples of the results of a query (the web documents) which are stored in folders.

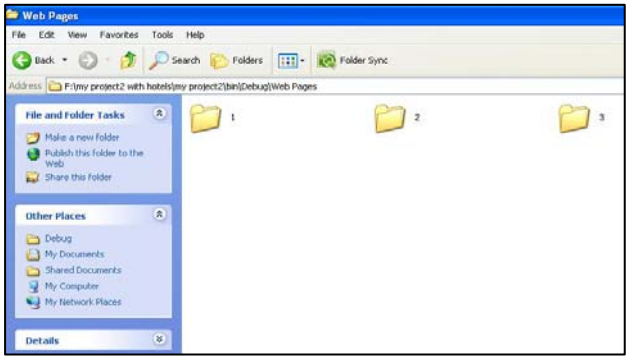

Figure 4. Examples of results of a query (web documents)

3.3 The Information Extraction (IE)

The IE extracts and classifies the web pages that are stored in the folders and converts them into text form. The details steps are discussed below.

Step 1: Based on the user-defined classification scheme of hotel domain such as General, Location, Category, Room, and Facilities (the attributes are shown in figure 5) which is stored in the UDCS database, the IE extracts the web pages. Figure 6 illustrates an example of the source code of a web page. The IE extracts only the texts which are found between the tag "<TABLE" and the tag "</TABLE>". Figure 7 shows an example of information which is extracted by IE based on the user-defined classification scheme of hotel domain. The IE ignores all texts which are not related to the user-defined classification scheme, that are used for programming HTML web pages such as cellspacing, TBODY, TR, TD, row, href, >, <, /, etc.

Attribute	Index_no
General	1
Location	2
Category	3
Room	4
Facilities	5

Figure 5. The user-define classification scheme related to hotels

```
<TABLE>
 <TBODY>
 <TR>
    <TD>Number of Rooms</TD>
    <TD>4</TD>
 </TR>
 <TR>
    <TD>Check-in time<TD>
    <TD>14:00 hrs</TD>
 </TR>
 <TR>
    <TD>Check-out time<TD>
    <TD>12:00 hrs</TD>
 </TR>
 <TR>
    <TD>Rating<TD>
    <TD>Two Stars</TD>
 </TR>
```

Figure 6. Example of the source code of a web page

```
- Number of Rooms
: 4

- Check-in time
: 14:00 hrs

- Check-out time
: 12:00 hrs

- Rating
: Two Stars
```

Figure 7. Sub attributes with their values saved in an array

Step 2: IE extracts the attributes, sub-attributes, and values of the sub-attributes and later identifies the index of attribute that is matched, as shown in Figure 8. IE classifies the extracted attributes and sub-attributes by grouping them based on the index number, as shown in Figure 9.

```
4- superior room
4: 170

4- family suite
4: 260

4- family suite quad sharing
4: 300

4- family suite quintet sharing
4: 380

4- junior suite
4: 320

4- aldy suite
4: 400
```

Figure 8. Attributes in text form

```
1* GENERAL
   - accommodation
      : To guarantee you amemorable stay with us,Aldy Hot

2* LOCATION
   - address
      : 27,JalanKota,Melaka

3* CATEGORY

4* ROOM
   - room facilities
      : AstroChannel,Coffeeandtea-makingfacilities,Indivi
   - rates are inclusive
      : Complimentary coffee/tea making facilities in roo
   - standard room
      : 150
   - superior room
      : 170
   - family suite
      : 260
   - family suite quad sharing
      : 300
   - family suite quintet sharing
      : 380
   - junior suite
      : 320
   - aldy suite
      : 400

5* FACILITIES
   - facilities
      : Aldy Hotel is famous for its happening dining and
```

Figure 9. Attributes after grouping together

Step 3: The IE converts the extracted and classified web page into text form. Figure 9 shows the web pages which are extracted, classified, and converted into text form by IE.

Step 4: The IE converts the text (web pages in text form) to structured information. The IE counts the number of attributes available in each text form and saves the results in a table. Figure 10 illustrates the number of attributes available in each text form.

Name of text	Number of extracted sub-attributes
Text 1	1
Text 2	12
Text 3	2

Figure 10. The structured information

3.4 The Relevant Information Analyzer (RIA)

The function of RIA is to determine the relevant information extracted from Information Extraction (IE) based on the number of attributes available in each text form. The steps performed by RIA are presented below.

Step 1: RIA receives the structured information from IE.

Step 2: RIA determines the relevant information extracted from IE based on the number of attributes in each text. For example, Text 1 has a single sub-attribute. RIA deletes Text 1 because it contains little information. Sometimes one text (web pages in text form) has the same information found in other text or lesser. In this case, RIA deletes one of the texts.

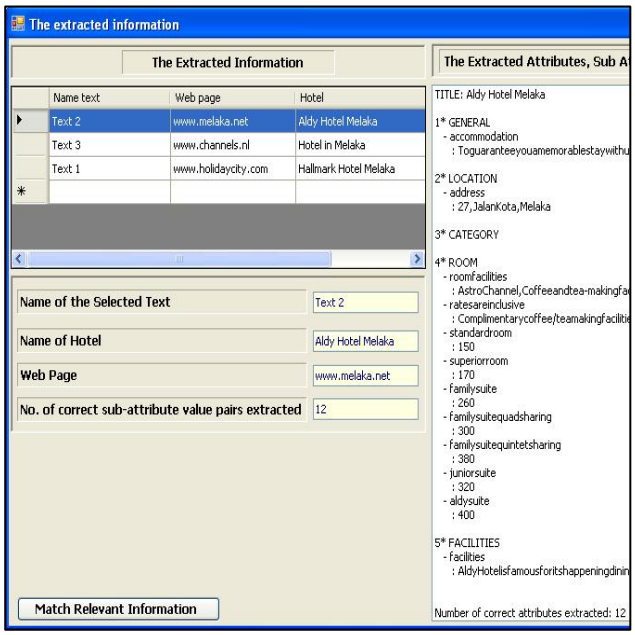

Figure 11. Results without user's context

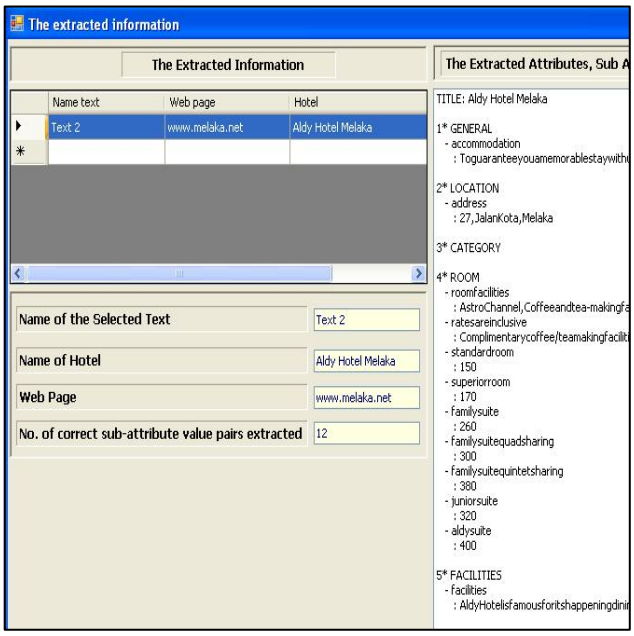

Figure 12. Results with user's context

Figure 11 illustrates the results which are displayed to user. At this stage no context has been considered. A user clicks on any texts in the form (Figure 11) then the web sites which have the information about the text appears in a list box.

3.5 The Match Relevant Information (MRI)

In order to retrieve relevant information that suits the user's preferences, the MRI performs a matching process which matches the user's preferences that are stored in the UPDB against the information that has been extracted by IE and later simplified by RIA. Only those texts whose extracted information matches the user's preferences will be selected while those texts that do not match are ignored. As an example, refer to Figure 11 which shows three texts that are relevant to the user's query but based on user's preferences only text 2 has been selected as shown in Figure 12. If more than one text that match the user's preferences, then those texts are ranked according to the number of matched sub-attributes/preferences.

Nevertheless, the matching process is a challenging task due to the following:

i. The sub-attributes extracted do not correlate directly with the user's preferences, i.e., identifying which extracted sub-attribute(s) should be considered to be compared to the user's preferences is not a trivial task.
ii. Some of the information related to user's preferences are not exact values but are range of values or approximation which further complicates the matching process.
iii. Even for a field of preferences, users might have several preferences. For example, user might prefer not only family suite but standard room or superior room, if the family suite is not available. Thus, identifying the priority of user's preferences is important so that the matching process is not limited to solution with only exact match.

4 Conclusion

In this paper, we proposed a framework to extract information from semi-structured web pages with user's context. The proposed framework provides facilities to the user during search. A user does not need to visit the homepages of companies to get the information. The information is extracted from the web pages and only the information that is relevant will be displayed to the user. Nevertheless, an information extracted by the IE might not be relevant if user preferences are not being considered. Thus, our proposed framework considers user preferences and only extracts the information that is relevant to user's preferences.

5 References

[1] Mahmoud Shaker, Hamidah Ibrahim, Aida Mustapha, and Lili Nurliyana Abdullah, "Information Extraction from Hypertext Mark-Up Language Web Pages", *Journal of Computer Science*, USA, p. 596-607, 2009.

[2] Anind K. Dey. "Understanding and using Context". *Journal of Personal and Ubiquitous Computing*, Vol. 5, Issue. 1, p. 4-7, 2001.

[3] Huanhuan Cao, Derek Hao Hu, Dou Shen, Daxin Jiang, Jian-Tao Sun, Enhong Chen, and Qiang Yang, "Context-Aware Query Classification", In *Proceedings of the 32rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, USA, p. 3-10, 2009.

[4] Junsong Yuan and Ying Wu, "Context-Aware Clustering", In *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement*, ACM Press, p. 217-230, 2008.

[5] Manli Zhu, Daqing Zhang, Jun Zhang, and Brian Y. Lim, “Context-Aware Information Display”, In *Proceedings of ICME*, IEEE, p. 324-327, 2007.

[6] Sung Won Jung, Kyung Hee Sung, Tae Won Park, and Hyuk Chul Kwon, "Intelligent Integration of Information on the Internet for Travelers on Demand", In *Proceedings of the ISIE, IEEE International Symposium*, Vol. 1, p. 338-342, 2001.

[7] Srinivas Vadrevu, Fatih Gelgi, and Hasan Davulcu, "Information Extraction from Web Pages using Presentation Regularities and Domain Knowledge", *Journal of World Wide Web*, Springer Netherlands, Arizona State University, Vol. 10, Issue. 2, p. 157-179, 2007.

[8] Horacio Saggion, Adam Funk, Diana Maynard, and Kalina Bontcheva, "Ontology-based Information Extraction for Business Intelligence", In *Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference*, Vol. 4825, p. 843-856, 2007.

[9] Gilles Nachouki, "A Method for Information Extraction from the Web", *Journal of Information and Communication Technologies*, ICTTA, Vol. 1, p. 517-521, 2006.

[10] Fei Hong and Zhuang Zhao, "Information Extraction System in Large-Scale Web", *Journal of Communications and Information Technology*, ISCIT, Vol. 2, p. 809-812, 2005.

[11] Veda C. Storey, Vijayan Sugumaran, and Andrew Burton-Jones, “The Role of User Profiles in Context-Aware Query Processing for the Semantic Web”, In *Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems*, UK, p. 51-63, 2004.

[12] Andrew Burton-Jones, Sandeep Purao, and Veda C. Storey, “Context-Aware Query Processing on the Semantic Web”, In *Proceedings of the 3rd International Conference on Information Systems*, 2002.

[13] Kostas Stefanidis, Evaggelia Pitoura, and Panos Vassiliadis, “A Context-Aware Preference Database System”, *International Journal of Pervasive Computing and Communications*, Vol. 3, No. 4, p. 439-460, 2007.

[14] Mohamed F. Mokbel and Justin J. Levandoski, “Toward Context and Preference-Aware Location-based Services”, In *Proceedings of the Eighth ACM International Workshop on Data Engineering for Wireless and Mobile Access*, p. 25-32, 2009.

[15] Arthur H. van Bunningen, Ling Feng, and Peter M. G. Apers, “A Context-Aware Preference Model for Database Querying in an Ambient Intelligent Environment”, In *Proceedings of the 17th International Conference on Database and Expert Systems Applications*, p. 33-43, 2006.

[16] Dana Al Kukhun, Bouchra Soukkarieh, Erick Lopez-Ornelas, and Florence Sedes, “LA-GPS: A Location-Aware Geographical Pervasive System”, In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering Workshop*, p. 160-163, 2008.