International Journal of Hybrid Information Technology
Vol.8, No.6 (2015), pp.115-124
http://dx.doi.org/10.14257/ijhit.2015.8.6.12

Interactive framework for Feasibility Studies using Data Mining

Sedeka Mahmoud El Shamy[1], Luo Jiawei[2] and Tamer Galal Moursi Farag[3]

[1,2,3]College of Computer and Communication, Hunan University, 410082
sedekaamahmoud@hotmail.com[1], luojiawei@hnu.edu.cn[2],
tamergalal@hotmail.com[3]

Abstract

Everyday internet users are collecting huge data, and the challenge they now facing is how to convert the raw data inside the relational databases to meaningful information for them to take effective decisions based on what is going on now and to predict what will happen in the near future. Data Mining and Web Data Mining deal with the extraction of interesting knowledge from the World Wide Web, the main usage of Data Mining is to discover the trend of business development and make the right decisions. This paper innovate an intelligence platform by presenting a case study to improve future Feasibility Studies. In our case we combined a lot of Data Mining techniques and Web Mining methods to product an interactive platform that can help users to get there goals.

Keywords: *Data Mining, Web Mining, Data Analytics, Feasibility Studies*

1. Introduction

The different types of data have to be managed and organized in such a way that can be accessed by different users efficiently. Therefore, the application of Data Mining techniques on the Web is now the focus of an increasing number of researchers. Several Data Mining methods are used to discover the hidden information in the Web. However, Web Mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified in such a way to better suit the demands of the Web. New approaches should be used to fit the properties of Web data in a better way. Furthermore, not only Data Mining algorithms, but also artificial intelligence, information retrieval and natural language processing techniques can be used efficiently. Thus, Web Mining has been developed into an autonomous research area. The rest of this paper is organized as follows. In Section 2, we present data mining detention and it is methods. In Section 3, the proposed approach is discussed in detail. Some cases studies are illustrated in Section 4 and Section 5 concludes this paper.

2. Background Work

2.1. Data Mining

Data Mining knowledge discovery is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data Mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data Mining tools can answer business questions that traditionally were consuming a lot of time to be resolved. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data Mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides [1].

2.1.1. Data Mining Methods : The commonly methods of data analysis by using Data Mining methods are including classification, regression analysis, clustering, association rules, characteristics, deviation analysis, *etc.*, [2]. They are mining data from a different perspective:

1) Classification: Classification is to find out the common characteristics, which are found from a set of data objects in database, and they can be divided into different classes in accordance with classification model. The purpose is mapped the data items of database to a given category through the classification model. It can be applied to the customer classification, customer attributes and characteristic analysis, customer satisfaction analysis, customer buying trends forecasting, *etc*.

2) Regression Analysis: Regression Analysis reflects the characteristics whose attribute value in transaction database temporal, generate a function that data item is mapped to real predictor variables and find the dependency relationship between variable or attribute. The main research issues include the trend characteristics of data sequence, the forecast of data series and the correlation between data. It can be applied to all aspects of marketing, such as customers seek, maintain and prevent activities of loss customers, product life cycle analysis, sales trend forecasting and targeted promotions, *etc*.

3) Clustering: Cluster analysis is dividing a set of data into several classes according to similarities and differences. The purpose of cluster analysis is making data similarity as large as possible between the same class and as small as possible between different classes of data. It can be applied to the classification of customer group, analysis of customer background, forecasting of customer buying trends, market segmentation, *etc*.

4) Association Rules: Association rules is the rule which describes exists relationship between data items in the database. According to the appearance of some items in the transaction, we can derive other items that also appear in the same transaction, and that is the association or interrelation hide in the data. In customer relationship management, we can find interesting association relationship from large number of records and critical factor who impact marketing effectiveness by mining mass data from enterprise customer database. It provides references for decision support which include orientation and price fixing of product, customize customer base, customers seeking, subdivision and maintaining, marketing and sales, marketing risk evaluation and fraud prediction *etc*.

5) Characteristic: Characteristic analysis is to extract characteristics formulas which express the general characteristics of the data set from a set of data in the database. Marketing personnel can get a series of reason and principal character which lead to loss customers by extract characteristic from the factors of loss customers, and it can effectively prevent the loss of customers in using characteristics.

6) Deviation Analysis: Deviation analysis includes a large class of potential interesting knowledge, such as the unusual instance in classification, the exception of pattern, the observed result on the expected deviation. The purpose is to seek the significant differences between the observed result and reference volume. Managers are more interest in those unexpected rules in business crisis management and pre-warning. Unexpected rule mining can be applied to the discovery, analysis, identification, evaluation and pre-warning of all kinds of abnormal information [3, 4].

2.2. Web Mining

Web Mining is the integration of information gathered by traditional data Mining methodologies and techniques with information gathered over the World Wide Web. The information gathered through Web Mining is evaluated (sometimes with the aid of software graphing applications) by using traditional Data Mining parameters such as clustering and classification, association, and examination of sequential patterns [5].

In general, Web Mining can be divided into three categories: Web Content Mining, Web Structure Mining and Web Usage Mining, see Figure 1.

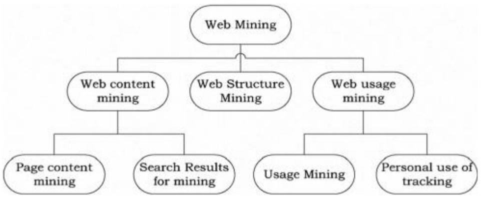


Figure 1. Shows the Classification of Web Mining

A) Web Content Mining extracts useful information or knowledge from Web page contents. For example, we can automatically classify and cluster Web pages according to their topics; we can also discover patterns in Web pages to extract useful data such as descriptions of products, postings of forums, *etc*., for many purposes. Furthermore, we can mine customer reviews and forum postings to discover consumer sentiments [6].

B) Web Structure Mining discovers useful knowledge from hyperlinks (or links for short), which represent the structure of the Web. For example, from the links we can discover important Web pages, which, incidentally is a key technology used in search engines. We can also discover communities of users who share common interests [7-10].

C) Web Usage Mining also known as Web log mining, its main objective is to find interesting model from the Web visit record. For the research in this area there are two main directions: 1) General access patterns track and personalized use record track and 2) Knowledge from Web page content.

3. Proposed Work

In this study, we provide information and advice to the users wishing to start their own projects in light of the available assets through making an interactive website. The users enter their data and assets and, by using the data mining process, the website provides a FS according to the provided data and other data that are provided by the website to the users to access to the available projects that fit their skills and experiences.

This project will cover the approaches of analysis and implementation of FS, we use SQL Server 2010 and visual studio 2012 to present this study. The proposed work is composed of four bottom-up layers: user layer, preparing layer, mining layer and decision layer. These four layers are closely linked, and all levels are included in Data Mining technology, aimed at dynamic interactive frame work to make feasibility study. The layers are shown in Figure 2.

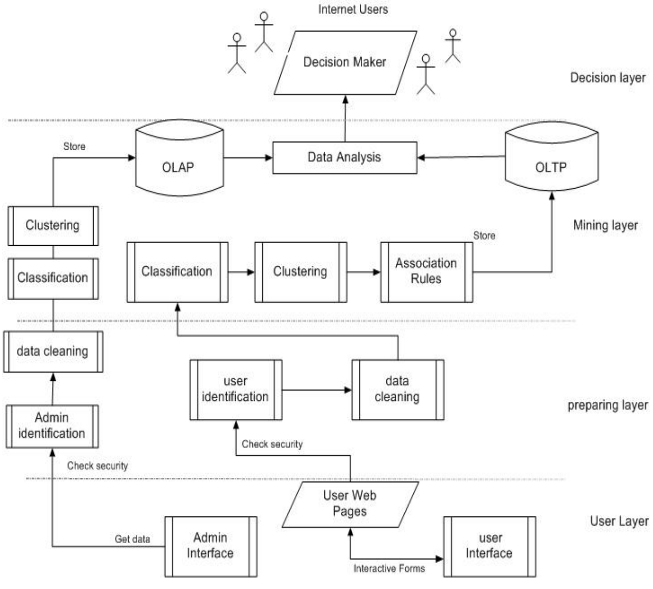


Figure 2. Shows the Proposed Four Bottom-Up Layers

3.1. User Layer

The task of User Layer is to collect Data Website forms and it’s converted two types:

1) User Interface; user web pages: They are interactive forms to get data from the user. We ask a question to the user and depending on the answer we ask another question. By this way, we guide the user to find what is best for him in order to finally be able to make the appropriate FS. For example, if we know his personal experience, qualities and so on, the system can automatically suggest to the user the best project that coheres with his abilities and skills.

2) Admin Interface: In this Section we collect data from different FS to provide the system with strong, accurate and fresh studies.

3.2. Preparing Layer

3.2.1. User Section

3.2.1.1. User Identification: Determining the single user must be done. The purpose of User Identification is to identify the users’ uniqueness. User Identification can be complete by using “User’ Registration Techniques” and “User Session Identification”. This should be done based on the User Identification with the aim of dividing each user's access information into several separate session processes. The simplest way is to use time-out estimation approach, that is, when the time interval between the page requests exceeds the given value, namely, that the user has started a new session.

3.2.1.2. Data Cleaning: Its main task is to remove the distortion from the forms. We use forms validation and JavaScript and function to handle entry data mistakes like spaces, special character, wrong or null values and so on, which are not associated with the useful data to narrow the scope of data objects for example using Trim()function to remove space from entered data and this script code to take only numbers

```
        <script type="text/javascript">
            function Getchar(event) { var chCode = ('charCode' in event) ? event.charCode :
event.keyCode;
                            //alert("The Unicode character code is: " + chCode);
                   return ((chCode >= 48 /* '0' */ && chCode <= 57 )/* '9' */ || chCode ==
46);}</script>
```

3.2.2. Admin Section: we use the same steps as User Section to identify the admin and after that clean the input data from the noise or distortion.

3.3. Mining Layer

The task of Mining Layer is to analysis and process the data by means of Data Mining models and algorithms, in order to mine out useful knowledge from our interactive website. This phase is core of our study. Classification, Clustering and Association rules can be used in our case study to analysis and prepare data to save them in OLTP database.

3.3.1. User Section

3.3.1.1. Classification: Classification aims to assign new data to different groups according to classification function or classification model. It can be used to classify users’ personal data and Feasibility Study Requirement data. This will be helpful to the formulation of Feasibility Studies based on users’ data and admin entered data. Classification analysis is also an important tool in Data Mining. There are some ways to achieve the classification analysis, such as “Decision Tree Classification” and “Neighboring Learning Method”, we use "Rule-Based Classification". Rules are a good way of representing information or bits of knowledge. A rule-based classifier uses a set of IF-THEN rules for classification.

3.3.1.2. Clustering: Clustering technique can divided the data items of the same feature into one group, so that the users' information with similar characteristics can be clustered out and analyze the user entered data, that can help the system to get a better understanding to their needs, to carry out the clustering operation. A two-step clustering algorithm can be used to target users aim with similar available Feasibility Studies in the same specialty. We use Rule-Based Classification in many different places in our code to present classification and clustering as follows:

```
  if (Convert.ToInt32(TxtMoney.Text.ToString()) >= 5000 &&
                    Convert.ToInt32(TxtMoney.Text.ToString()) <= 10000 && haveplace.Checked == true)
      {  dt = dbphd.ExecSp("ForWeb", new SqlParameter("@DMLType", "GetAllProj"),
                                      new SqlParameter("@specialty", Wspecialty.SelectedValue),
                                      new SqlParameter("@experience",
Workexperience.SelectedValue));
        if (dt != null && dt.Rows.Count > 0) {
              Label1.Text = "available projects:";
                    for (int i = 0; i < dt.Rows.Count; i++)
                       {  Label2.Text = dt.Rows[i]["PrjName"].ToString(); } }
        else
        {  Label1.Text = "No projects found";
           Label2.Text = "";  }
```

3.3.1.3. Association Rules: The purpose of Association Rule is to uncover the hidden relationships among the data. Association Rules are used for mining out the related rules from the sequence items in user access sequence data. In our case, we analyze personal data information like age, education, living place, experience and skills to find the Association Rules in order to get valuable information from the given data so that the system can present suitable Feasibility Studies, if age "20" to "30" and education is "Average" so user take handmade projects .

3.3.2. Admin Section: We use the Classification and Clustering as in User Section to analysis the entered data by Admin Section we add about 50 FS as a base for our work contain different projects and budgets to service the desires of investors and after that saving data in OLAP, see Figure 3.

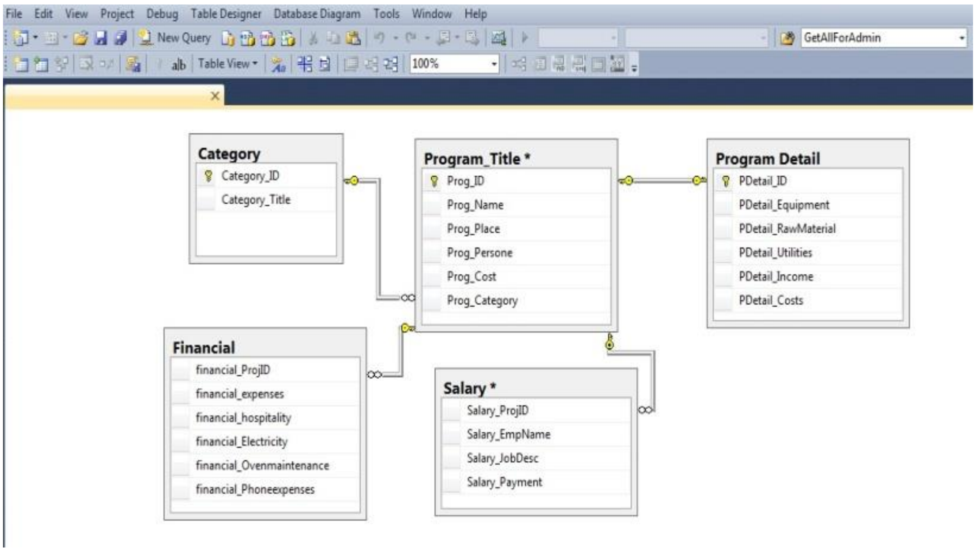


Figure 3. Shows the Admin Section Tables

3.4. Decision Layer

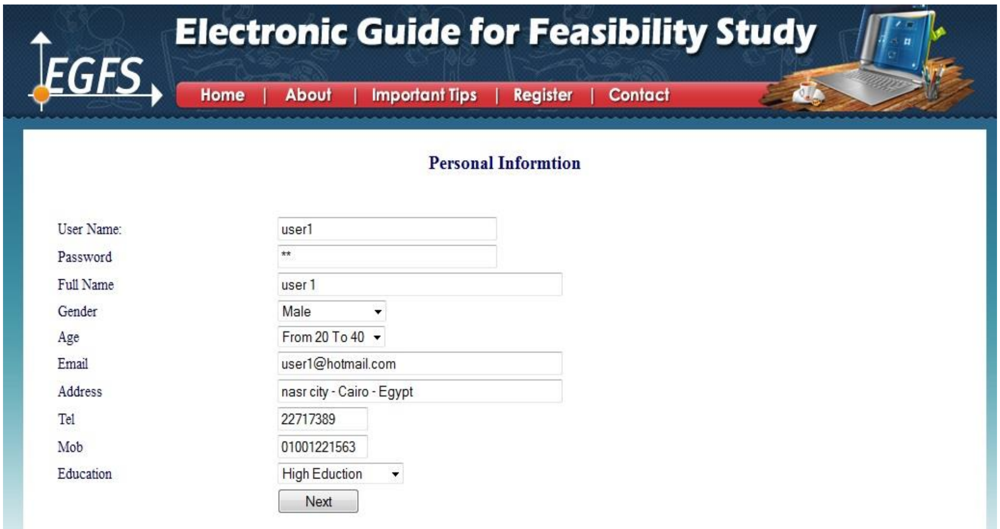
Decision Layer is a key part in our case study. Its main task is to make interactive website data (OLTP) combined with the Data retrieved and saved from Admin Section (OLAP) according to the knowledge base established in Mining Layer to present all available Feasibility Studies according to his skills and experience.

4. Model Application “Some Cases as an Example in This Section”

We will take some example for our website to explain, demonstrate and analyze our work in real cases.

4.1. Case No. 1

User fills two forms as follow: the first one contains personal information. The user adds all his data like his name and password. The user will use it to login after register to add his achievement and feedback about his work. For our analysis we use gender, age, address and education to classify the users according to these parameters as see Figure 4.

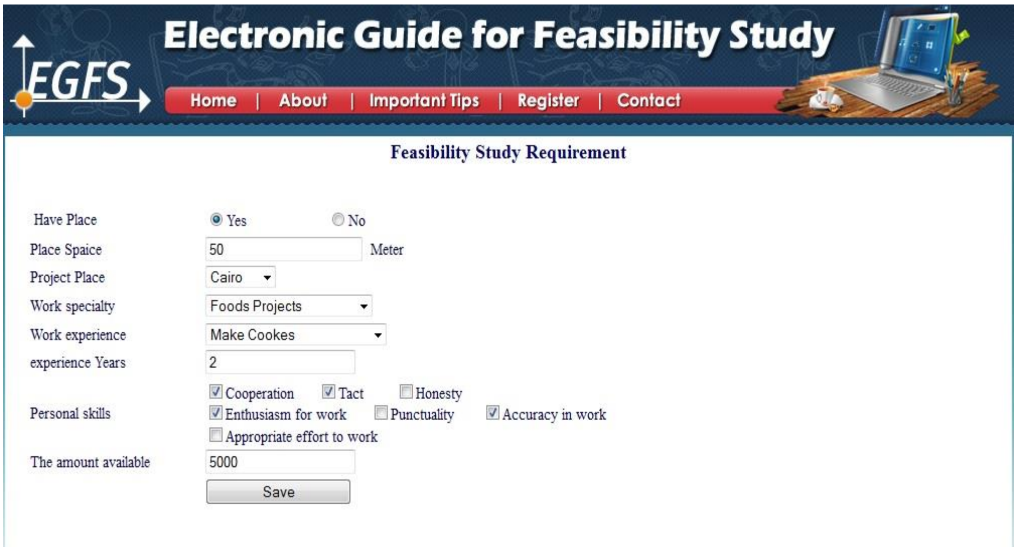

EGFS
Electronic Guide for Feasibility Study
Home | About | Important Tips | Register | Contact
Next

Personal Informtion

User Name:	user1
Password	**
Full Name	user 1
Gender	Male
Age	From 20 To 40
Email	user1@hotmail.com
Address	nasr city - Cairo - Egypt
Tel	22717389
Mob	01001221563
Education	High Eduction

Figure 4. Shows the Personal Information Form

In the second one the user adds more detailed data about his requirement. First, he is asked “Have a place?”, “Yes” or “No”, if “Yes” he will be asked about “Project place?” and “Project space in meters?” After that we ask him about “Work specialty?” and we feed this item from the Admin Section with all the available specialties we have in our database, in this case “food project”. The next question is about “Work experience?” and is also filled from the Admin Section depending on the previous choice from “Work specialty”, in this case “making cookies”. Then he is asked about “Experience years?”, in this case “2 years”. After that “Personal skills?” and finally “How much money he has to start the project?” in our case “10000”, see Figure 5.


Electronic Guide for Feasibility Study
EGFS
Home | About | Important Tips | Register | Contact

Feasibility Study Requirement

Have Place	● Yes ○ No
Place Spaice	50 Meter
Project Place	Cairo
Work specialty	Foods Projects
Work experience	Make Cookes
experience Years	2
Personal skills	☑ Cooperation ☑ Tact □ Honesty ☑ Enthusiasm for work □ Punctuality ☑ Accuracy in work □ Appropriate effort to work
The amount available	5000

Figure 5. Shows the Feasibility Study Requirement Form

At the end the system combines the data entered by the user and the data entered by the admin and uses our proposed Data Mining techniques to offer to the user all the available Feasibility Studies to choose the suitable one for him. In this case, the system presents two choices":

1- Cookies project 2- Pies project

The user chooses "cookies project". The FS for that project is as follows:

- Project Name : Making Cookies
- Project Idea: Making home cookies for events and shops.
- Project characteristics and inputs, see Table 1.

Table 1. Shows Project Characteristics and Inputs

Machinery and equipment	Oven and Hand Tools
Place	Home
Workforce	Employer and an assistant
Raw material (goods)	Flour, Yeast, Oil, Powdered milk, Eggs, Packaging materials
Utilities	Water and Electricity

- Project output: Cookies
- Most important financial characteristics see Table 2.
-

Table 2. Shows Most Important Financial Characteristics

Project cost	10558
Working capital	8658
Fixed assets cost	900
Incorporation expenses	1000
Return rate on investment	-91
Expected profit in 1st year	-0.86%

- Capital payback period, see Figure 6.
-

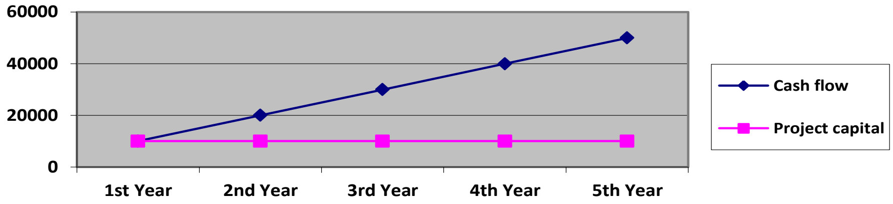


Figure 6. Shows Capital Payback Period

- List of revenues and costs, see Table 3.

Table 3. Shows Revenues and Costs

Product/Service	Monthly quantity	Unit value	Monthly revenues	Annual revenues
Circular cookies (53g)20 pieces	500	10		
Total revenue			5000	60000

- Wages and salaries, see Table 4.

Table 4. Shows Wages and Salaries

Job	Number	Monthly salary	Monthly salaries	Annual salaries
Employer	1	1500		
Assistant\ assistant in sales and distribution	1	1000		
Total			2500	30000

- Administrative expenses, see Table 5.

Table 5. Shows Administrative Expenses

Item	Annual costs	Monthly costs	Notes
Public relations and hospitality	500	42	
Phone expenses	1800	150	
Total	2300	192	

- Utility costs and energy, see Table 6.

Table 6 Shows Utility Costs and Energy

Item	Monthly costs	Annual costs
Electricity	250	3000
Total	250	3000

- Costs of maintenance and spare parts, see Table 7.

Table 7. Shows Costs of Maintenance and Spare Parts

Item	Annual costs	Monthly costs
Oven maintenance	240	20
Total	240	20

- Operating costs excluding depreciation

Table 7. Shows Operating Costs Excluding Depreciation

Item	Monthly costs	Annual costs
Raw material coast (or purchases value)	474	5688
Wages and salaries	2500	30000
Administrative expenses	192	2300
Utilities and energy costs	250	3000
Maintenance and spare parts costs	20	240
Total	3436	41228
5% reserve from total operating costs	172	2061
Total operating costs + reserve	3607	43289

4.2. Case No. 2

The user fills the forms as we mentioned before. We entered data related to gender, age, address, education, project place and space and personal skills. Regarding “Work specialty” his choice was Art Projects, “Work experience”: Handmade, “Experience years”: 4 years, “How much money he has to start the project?” 7000 and the system presents these projects:

1- Candle project 2- Accessory project 3- Toys project

This Feasibility Study is just an example and the other cases will follow the same steps as we mentioned before.

5. Conclusion

In this paper we present an interactive framework to help investors to make a good Feasibility Study by taking their requirement and translated it to real project. By experiment we demonstrate that our proposed framework meets the user’s requirement for Feasibility Studies, while, in the future work we will continue our work in the same felid of interactive Web Data Mining (content, usage) to improve the project condition after the user began his real project to get high income with good quality.

References

[1] L. Xinying and W. Peizhi, “Data Mining Technology and its Application in Electronic Commerce”, Wireless Communications, Networking and Mobile Computing, WiCOM'08, 4th International Conference on, IEEE, **(2008)**.

[2] Y. Chen and F. Wang, “A Dynamic Pricing Model for E-commerce Based on Data Mining”, 2009 Second International Symposium on Computational Intelligence and Design, vol. 1, **(2009)**.

[3] R. Kohavi and F. Provost, “Applications of data mining to electronic commerce”, Springer US, **(2001)**.

[4] S. Bin, L. Yuan and W. Xiaoyi, “Research on data mining models for the internet of things”, Image Analysis and Signal Processing (IASP), International Conference on, IEEE, **(2010)**.

[5] G. Yu, C. Xia and X. Guo, “Research on Web Data Mining and its Application in Electronic Commerce”, Computational Intelligence and Software Engineering, CiSE, International Conference on. IEEE, **(2009)**.

[6] L. Mei and F. Cheng, “Overview of Web mining technology and its application in e-commerce”, Computer Engineering and Technology (ICCET), 2nd International Conference on IEEE, vol. 7, **(2010)**.

[7] X. Huang, D. Chen and Q. Zu, “Research and application of Web mining based on Web service”, Pervasive Computing and Applications (ICPCA), 2010 5th International Conference on, IEEE, **(2010)**.

[8] R. Zhao and S. Cao, “The Study of Users and Content Comparison Based on Multiple Chinese Microblogs”, Computational Intelligence and Design (ISCID), 2012 Fifth International Symposium on IEEE, vol. 1, **(2012)**.

[9] O. T. Ali, A. B. Nassif and L. F. Capretz, “Business intelligence solutions in healthcare a case study: Transforming OLTP system to BI solution”, Communications and Information Technology (ICCIT), Third International Conference on IEEE, **(2013)**.

[10] S. Zihao and W. Hui, “Research on recommender system in e-commerce based on Web mining”, Advanced Management Science (ICAMS), International Conference on IEEE, vol. 2, **(2010)**.

Authors

Sedeka Mahmoud El Shamy, is a student in Hunan University, in the College of Computer and Communication, she is PhD student, and work in the topic of Data mining and her email sedekaamahmoud@hotmail.com.

Luo Jiawei, is a vice dean and professor in the College of Computer and Communication, Hunan University and her email luojiawei@hnu.edu.cn

Tamer Galal Moursi Farag, is a student in Hunan University, in the College of Computer and Communication, He is PhD student, and his email tamergalal@hotmail.com.