Implementation of Apriori Algorithm to Analyze Organization Data: Building Decision Support System

Abdullah Saad Al-Malaise
Information System Department
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah
Kingdom of Saudi Arabia.

ABSTRACT
Building decision support system is major concern for almost every organization to get decisions on daily processes. In current market situation automated decision support systems can produce more alternatives (multi criterion) for decision makers. In this paper we propose automated decision support system with integration of data mining techniques. Building system with amalgamation of both techniques showing feasible approach that can produce appropriate results and fast processing. In presented model the data mining (DM) abstract will support to generate new rules and patterns on customers/employees queries and data. Whereas decision support system (DSS) abstract can ask help from DM databases online or offline to provide multi criterion alternatives. The main purpose of this model is to help decision makers by using multi criteria decision making strategy with DM techniques as those consider powerful tool for decision making processes. In the end we have provided practical implementation using some real world data, to show step by step meaningful purpose of the proposed model.

Keywords
Decision Models, Association Mining, Knowledge Management.

1. INTRODUCTION
Decision support systems (DSS) has several building phases such as intelligent, design, choice and implementation as discussed [2]. In this scenario the extension is always available in building DSS. Because all of the four phases cover vast amount of strategies and building techniques. Keeping this in mind in this research we have integrated data mining phase as a essential part with DSS previous four phases. While data mining is use for extract novel information from the large databases. In the current scenario DSS database must contain decision rules, decision models, and knowledge management (an optional phase of DSS). So it shows that data mining can impact highly on DSS database.

Large databases are essential components for each and every organization survival. It will help the organization to initiate new projects and to take right decisions. These databases have more complex data for the implementation of any task because of its multi dimensional attributes. The different type of format and attributes of data create more difficulties includes text, images, videos, graphs etc.
This paper is an extended version of our previously published paper [13]. In this extension we showed some enhancement in our previous model. In addition we implemented the model by using data and tool discusses in the later case study sections. Eventually, the main purpose of this research is to reduce pressure from the decision makers to use same purpose tools and techniques together. In the methodology sections we defined it in detail. Before moving to the model, presented small introduction of DSS and data mining approaches in the succeeding sections for more clear understanding of the model.

1.1 An overview on Decision Support System
DSS consider an special type of information system which support decision maker to take decision fast, reliable and accurately. A DSS has its unlimited features, adding more techniques to it will provide more fruitful results. Organizations want to build DSS for making their decision practical and anticipative. The development of DSS can be based on data or model. According to Turban, the major classification of DSS based on (i) Data Oriented DSS, (ii) Model Oriented DSS [2]. Whereas there are several other extension has been presented from different scholars [13, 14].

Currently, the businesses running in the market are more complex and competitive. There are always different types of pressures come to decision makers which makes working environment more enthusiastic and hard working. Since, DSS is the system which help you deal with any pressure anytime. It makes decision makers to take more easy and timely decisions. Since Computerized support will help the decision makers to deal with organization, customers, employees, and competitors at the same time. Finally
in this complex environment the best decision can lead the organization on top.
DSS is one of the main tool deal with several type of business environmental factors such as; markets, customer demand, technology, and societal factors for generating best decision from the list of alternatives. A DSS has many phases, whereas each phase has several sub tasks to analyze, formulate and finally compose the decision. As Turban described that a DSS must consider four phases for complete decision making process such as; Intelligent, Design, Choice and Implementation [2].

[14], also outlined the broad view of DSS components in figure-1 [14]. Although in this figure the part of DSS includes all the sub-parts of DSS as described above. Figure-1 illustrates the large view of DSS, while keeping this in mind we proposed an enhanced version of the model in the methodology section, with contribution of data mining strategies to make decision making more powerful and full of options.

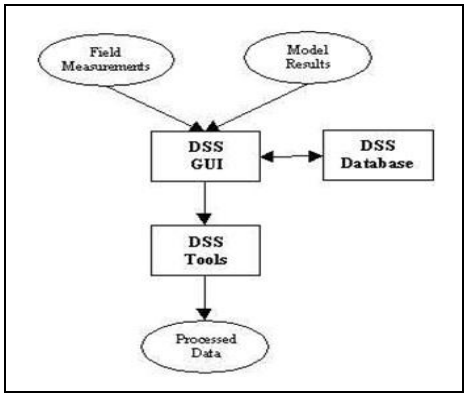

Fig 1: Components of DSS [14]

Before discussion on methodology, in the subsequent section we discuss data mining approaches and techniques to make this research more clear and understandable.

1.2 An overview on Data Mining Techniques

Data Mining, also known as knowledge extraction from large databases, where extracted knowledge consider nontrivial, unformulated and unknown but useful information. Data mining and knowledge extraction are frequently uses as synonyms. In fact data mining is a process to extract unknown and novel information from large databases which has much detailed and concise procedure. [5, 8]. In addition, Abdullah et.al explained that data mining implementation has major impact in building of DSS. other than the four major phases of DSS, data mining techniques also consider as one of the phase which can help the decision makers to dig out the new information from decision process data. Generation of the new rules and decision are the major resulting factors for data mining integration with DSS [9, 7].

Mainly, data mining tasks has been divided into descriptive and predictive methods. Classification, clustering and rule association mining are most common techniques use for predictive and descriptive analysis [10]. Therefore, mainly scholars describe data mining in three major tasks. As Zaine [5] stated in his book chapter about major techniques of data mining as follows:

a) Classification: Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification consider as an important task of data mining. Using this approach data must be already defined a class label (target) attribute. Firstly we divide the classified data into two sets; training and testing data [11]. Where each datasets contains others atrributes also but one of the attributed must be defined as class lable attribute. Jiawei Han [11] described classification task in two steps process; first is model construction and the second is model usage. The main target of this task is to build the model by using training dataset and then assign unseen records into a class by using the trained model as accurately as possible. While training data set is use to build the model on the other hand testing data set is use to validate the model [10].

b) Clustering: Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification.

Clustering is one of the major task has been applying for data mining, work on unsupervised data (no predefined classes) [12]. Clustering is a collection of data objects, clustered by taking similar object to one another within the same cluster, and dissimilar to the objects related in other clusters. Cluster differentiate by using similarities between data according to the characteristics found in the data and grouping similar data objects into clusters [11].

c) Association: Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets.

Data can be use to find association between several attributes, generate rules from data sets, this task is known as association rule mining [12]. Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction. The goal of association rule mining is to find all rules having support ≥ minsup (minimum support) threshold and confidence ≥ minconf (minimum confidence) threshold [10].

Moreover, association rule mining can be viewed as a two-step process, first, find all frequent itemsets: items satisfying minimum support. Second, generate strong association rules from the frequent itemsets: these rules must satisfy minimum support and minimum confidence [11].

2. METHODOLOGY

In Figure-2, presents the common data mining process proposed by [15]. It has limited step to apply any data mining technique on different kind of data. Whereas every step can have several subtasks to perform properly. Data mining approach always start from identification of the problem according to the selected data. Then data preprocessing such as; transformation, normalization, missing values and etc are some crucial task to apply before implementation of data mining technique. As without improper and wrong data selection may lead to the false result and expectations. Moreover [1, 6] also described that other than these basic steps of data mining cycle there can be several other sub-processes to perform accordingly.

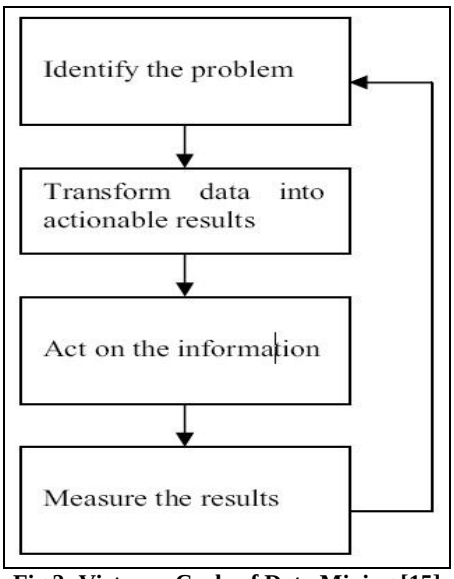

Fig 2: Virtuous Cycle of Data Mining [15]

Reference to the figure-1 and figure-2, which showed broad view of building DSS and DM processes respectively. In the subsequent section we proposed an integrated and enhanced model to combine DSS and DM process both in single system. The enhancement represents the micro view of both techniques. In addition we have applied this model in real world data to describe the importance and ultimately the generation of rules and patterns, showed in the final result section.

2.1 Proposed Model Description – Figure-3

The major contribution in this research is the proposed model illustrated in figure-3 which actually referenced from our previous published paper [13]. A novel approach to build DSS in combination of data mining technique to get better results from large database. Generation of the new rules, decision model and finally the optimal decision is the main cause of the presented model. For more understanding the set of generated rules and patterns presented in the results later section. On the whole the proposed model can split into two major section; DSS and DM.

DSS considers first part of the model; the idea is adopted from [14] that illustrated the basic components of DSS. It has front end graphical user interface which is directly connected with user interface. Receiving the query, finding the best option from its own database or contact with data mining interface if not find any answer from DSS database. Data mining abstract will timely maintain its database known as knowledge management. Knowledge management will keep all queries and their replied answer in it. This shows the general working of DSS abstract. Although DSS has several techniques to apply as per requirements such as; Analytical hierarchal process (AHP) [16], and what-if analysis [17].

The main reason for direct connection of DSS and knowledge management with query facilitator is to look first for the asked query in the local database or knowledge management. If the query needs some additional manipulation then it will be asked from DM abstract to generate new patterns for the newly arrived query. DSS tool can also be asked to apply the suitable DSS technique for the particular query. On the whole, majorly we need here to generate list of alternatives or choices which is the basic purpose of DSS. The list of alternatives with some pros and cons can help to select optimal decision. Final decision will always update in the knowledge management for the future references and forecasting purpose.

DM is the second major part in the model has many tasks to perform simultaneously. As discussed previously data mining abstract will invoke only when DSS interface would like to perform some extra strategy on the data. In fact it is helpful to connect such kind of system with data mining component, indeed DM is powerful tool to work on large operational databases. DM always start from the data preprocessing, it need to classify data, clean noise from the data and finally data transformation if needed. Then it has some common tasks to perform according to the requirement and structure of the data such as; Clustering, Association and Classification. In our case study presented in the later section we have applied association technique (Apriori) to generate rules from previous customer queries.

In the end, all applied queries, decision models, generated rules update in the knowledge management for the future reference and correspondence with customers. The same experienced data can also available for decision makers to analyze, forecast or take decision for future perspective. To know the actual working of the model we applied this model using case study of customer queries in the succeeding section. Screenshots of different stages make more clear understanding and practical working of the model.

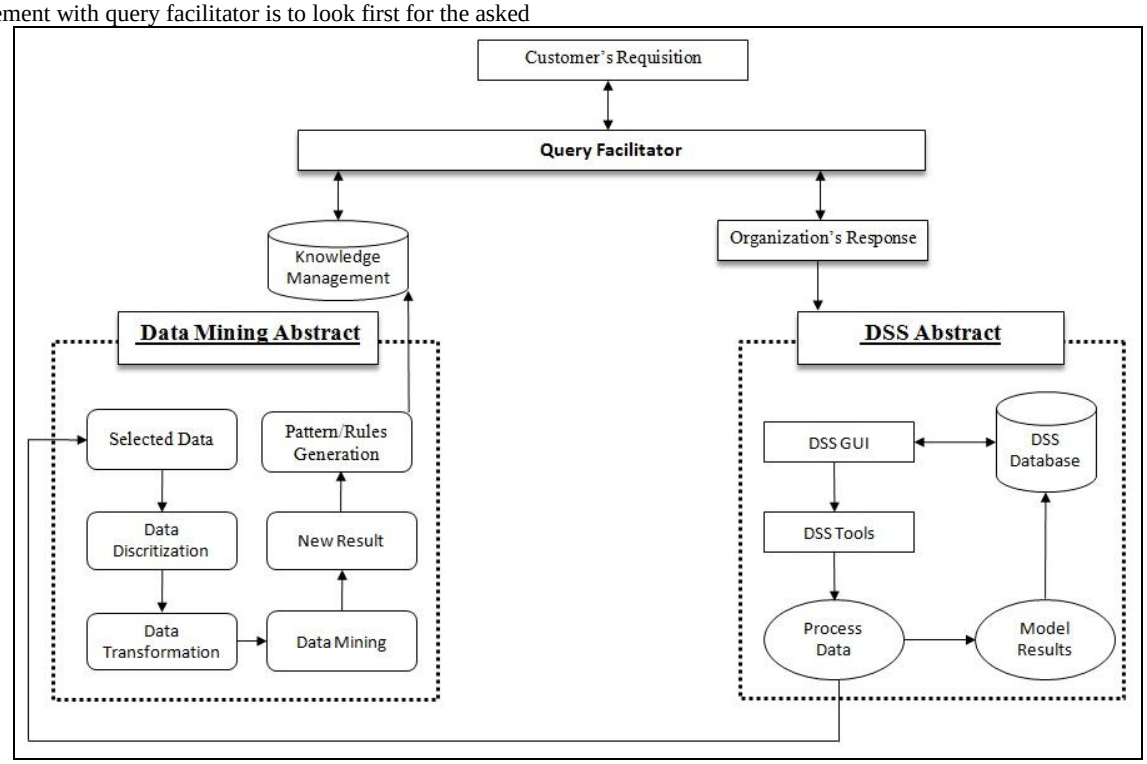

Fig 3: Proposed and Enhanced Model of DSS with integration of DM View (Compiled by Author)

2.2 Case Study of Customer's Queries

In this section, we applied our model (data mining part) using customers/employees queries data. On demand of the company to keep organization's information secret, we are not mentioning the name of the organization in our research. But this enterprise is doing business of home accessories dealing with different kind of customers and employees daily. They are dealing with customer/employees online and offline on different kind of queries such as; purchasing and selling of the goods, complaints, and etc. We arranged the data file as our requirements, because we used association technique to generate rules from the customer queries file to solve same kind of issues automatically in the future.

For this, we selected Apriori algorithm (association mining) to apply on data. Apriori is a famous algorithm use in association mining to generate learning association rules. It is build to operate on data file having transactions such as; collection of queries, item purchased together, etc. [3, 4].

2.3 VB Interface and Code

There are several tools available to build any data mining process. But here in this research we design a small program using Visual Basic programming language specifically to build Apriori Algorithm (Association Mining). VB interface is use for selection of input file and generate rules and patterns as a result or output file. Figure-4 & 5 displayed the interface and code window of this program respectively.

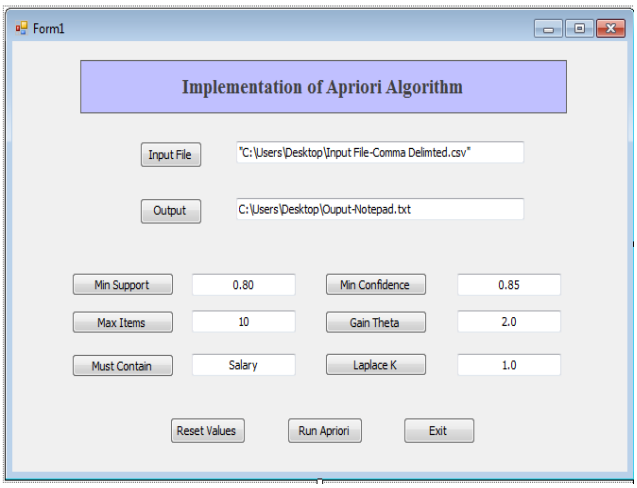

Fig 4: Apriori Algorithm Implementation – Interface Window

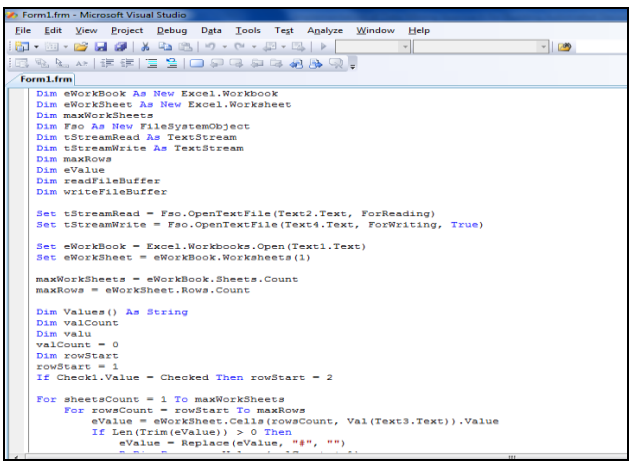

Fig 5: Apriori Algorithm Implementation - Code Window

2.4 Input Data File

Input file showed in figure-6, is selected data according to the association mining requirements. We have already applied the data transformation to transfer the data from the database format to excel (column) format. Then we also applied transformation of the data from text into binary format. Attribute (Column) selection, discritization and data cleaning techniques (replace missing value, outlier treatment) has also been applied to make proper data file before implementation. For this example the details of given input parameters can be seen in Figure-4, such as Support and Confidence value.

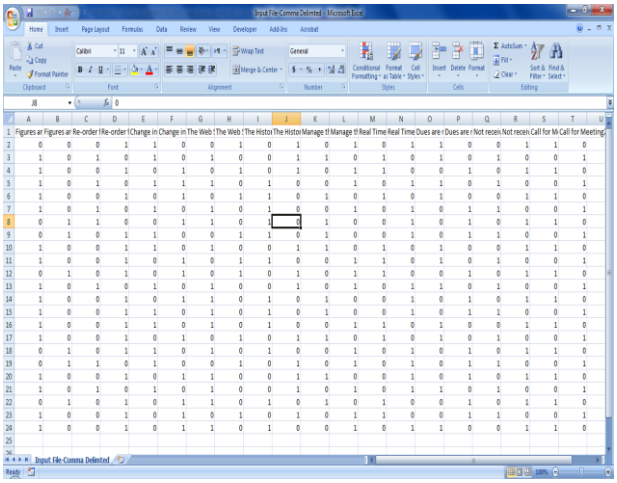

Fig 6: Input File – Excel (Column) Format

2.5 Output Generated File

In figure-7 presented the output generation. The format we selected for output generation is notepad for easy understanding and readable format. This output is based on the support and other parameters given at the runtime. These all values and format of the output file is changeable, it can be change according to the requirements. This file showed the all items those are under the support value given in the table. It display the items based on their repetition together as Apriori algorithm work also like this.

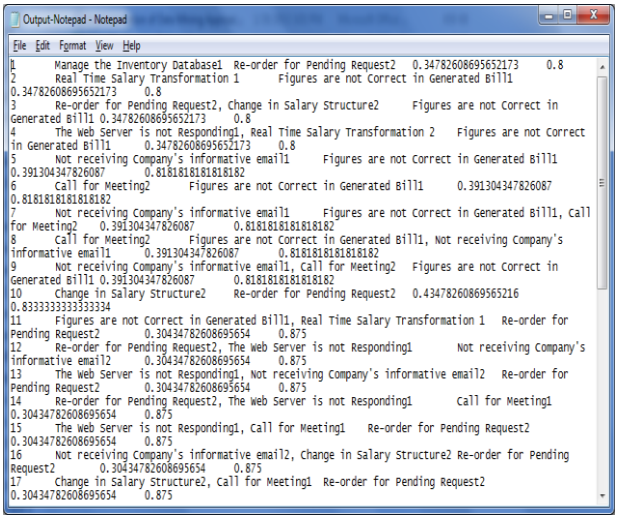

Fig 7: Output File – Notepad (Text) Format

2.6 Results

Finally, Figure-8 has been generated which displayed the rules extracted from the output file after given support and

confidence value. Altogether there are many rules generated here it can be more or less based on confidence value. The rules are working like if-then statement i.e. if salary is not transferred→web servers is not responding. This example results are based on apriori algorithm, changing in the data file and data mining technique can also give some suitable results.

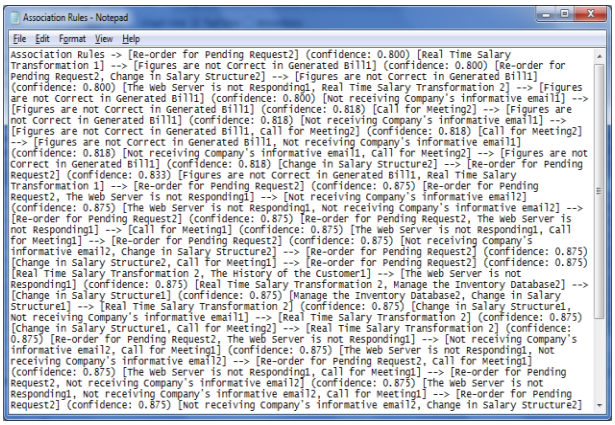

Fig 8: Result File – Rules and Patterns Generation using Support and Confidence Value

3. CONCLUSION

Competitive market demands in these days to analyze organization's database more frequently. As data mining consider powerful tool to analyze the data and it has the capability to merge with any system. Therefore we proposed and applied model of DSS with integration of data mining tasks. The practical implementation of the model can be helpful for the decision makers to modify and adapt according to their organization. Result shows the powerful working of the proposed model where rules and pattern generated using customer data are the main output of this research. In future, working on some other kind of databases can highlight the customize quality of the proposed model. This model has the capabilities to adapt by any organization as now it is in general customized form.

4. REFERENCES

[1] Abdullah Saad Almalaise Alghamdi, "Efficient Implementation of FP Growth Algorithm Data Mining on Medical Data", International Journal of Computer Science and Network Security (IJCSNS)-2011.

[2] Efraim Turban, Ramesh Sharda, Dursun Delen, Decision Support and Business Intelligence, 9th Edition, published by Pearson Education, Prentice Hall-2011.

[3] Apriori Algorithm, http://en.wikipedia.Org /wiki/Apriori algorithm #Algorithm, Last accessed Date: 1st January, 2013.

[4] Agrawal R, Imielinski T, Swami AN. "Mining Association Rules between Sets of Items in Large Databases." SIGMOD. June 1993, 22(2):207-16.

[5] Osmar R. Zaïane, "Chapter I: Introduction to Data Mining", CMPUT690 Principles of Knowledge Discovery in Databases, 1999.

[17] http://office.microsoft.com/en-001/excel-help/ introduction-to-what-if-analysis-HA010243164.aspx, last accessed date, 15th January, 2013.

[6] Abdullah Al- Mudimigh, Farrukh Saleem, Zahid Ullah, The Effects Of Data Mining In ERP-CRM Model – A Case Study Of MADAR, WSEAS Transaction, 2009

[7] Dr. Abdullah Al- Mudimigh, Farrukh Saleem, Zahid Ullah, The Role Of Data Mining In ERP-CRM Model", International Conference on Applied Computer & Applied Computational Science (ACACOS '09), Hangzhou, China, 2009.

[8] Data Mining for Customer Queries in ERP Model- A Case Study of MADAR", IEEE / ICICT-2009, Third International Conference on Information & Communication Technologies, Karachi, Pakistan, 2009.

[9] Dr. Abdullah Al- Mudimigh, Farrukh Saleem, Zahid Ullah, Efficient Implementation of Data Mining: Improve Customer's Behavior", The 7th IEEE/ACS, International Conference on Computer Systems and Applications, Rabat, Morocco, to be held on May, 10-13, 2009.

[10] Pang-Ning Tan, Michael Steinbach & Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2005, ISBN 0321321367

[11] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques", 2nd edition, 2005, Morgan Kaufmann, ISBN 1558609016

[12] Farrukh Saleem, Areej Malibari, Data Mining Course in Information System Department–Case Study of King Abdulaziz University, IEEE, ICEED- 3rd International Congress on Engineering Education, 7th – 8th December, 2011.

[13] Farrukh Saleem, Abdullah Al-Malaise, Implementation of Data Mining Approach for Building Automated Decision Support Systems, International Conference on Information Society (i-Society 2012), IEEE, June 25-28, 2012, London, UK.

[14] Fernando Toro and Roberto Mayerle, A Decision Support System for Enhancing Model Development and Application, Esri International Conference-2001.

[15] Hai Wang, Shouhong Wang, Medical Knowledge Acquisition through Data Mining, Proceedings of 2008 IEEE International Symposium on IT in Medicine and Education.

[16] T. L. Saaty. Relative Measurement and Its Generalization in Decision Making. Why Pairwise Comparisons are Central in Mathematics for the Measurement of Intangible Factors. The Analytic Hierarchy/Network Process. Rev. R. Acad. Cien. Serie A. Mat.VOL. 102 (2), 2008, pp. 251–318