# *Predicting user entries by using data mining algorithms*

Basel A. Alhaj
Faculty of Information Technology
Islamic University of Gaza
Gaza, Palestine
baselhaj@gmail.com

Ashraf Y. A. Maghari
Faculty of Information Technology
Islamic University of Gaza
Gaza, Palestine
amaghari@iugaza.edu.ps

***Abstract*—The information systems are widely spread in most official institutions, and become certified in all areas of our life such as education, health and entertainment. Usability is one of the most important factors, which encourages users to deal with these systems or refuse it. Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. In this paper we analyze the stored data in database of Palestinian Government decisions system in order to study the relationship between some attributes. Accordingly, we can find patterns that help us to make the system more user-friendly by offering suggestions to the users during data entry process. Naive Bayes, Rule Induction, K-NN, and Decision Tree methods are applied to the stored data in order to produce a prediction model that predicts entries to the user during the entry process, which can make the entry system more user-friendly. The experiment result shows the Naïve Bayes is the best model among the other techniques by achieving the highest accuracy of 68.41%. Future efforts can apply this model in the Government decisions system of Palestinian Ministers Council in Gaza.**

***Index Terms — Usability; Data Mining; Recommender Systems; Entry Prediction.***

## I. INTRODUCTION

Recent years have seen wide spread of the use of information systems in many of the official institutions and these systems become certified in all areas of our life such as education, health and entertainment. With the spread of these systems the issue of usability as a measure of the quality of these systems has emerged [1].

Usability defined by [2] as "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use". Usability is one of the most important factors, which encourages users to deal with these systems or refuse it. And eventually lead to the success or failure of the system [3].

Databases for these systems contain a large volume of data that is utilized only in the storage and retrieval of data. Data mining is the process of analyzing data into useful information. This process is done for finding correlations or patterns among dozens of fields in large relational databases [4].

Many researchers introduced works in software usability evaluation using data mining and other researchers proposed data mining methods for recommender systems.

In this paper we study the relationship between some attributes from the database of Palestinian Government decision system in an attempt to design prediction model that help us to make the system more user-friendly by suggesting entries to the system user during the entry process. Instead of selecting required entries from a large number of items, the prediction model is proposed to suggest related items from which the user can easily select entries.

The rest of this paper is organized as follows: Section two is background of data mining and classification methods Section three review of related works. Section four presents the methodology of this research. Section five explains results and discussion. Finally section six presents research conclusion.

## II. BACKGROUND

Data mining defined as "Is the process of extracting knowledge hidden from large volumes of raw data. The knowledge must be new, not obvious, and one must be able to use it" [10].

Data mining tasks classified into two main categories: descriptive (unsupervised learning) and predictive (supervised learning). The tasks of descriptive data mining characterize the general properties of data but predictive data mining tasks perform processes on the data in order to make predictions though finding patterns [10].

Classification is predictive supervised learning task. It divided into two processes. The first process is find patterns to build the model depending on the class attribute of the data set. Then the second process is test the designed model using new unseen data usually part of the whole data set [11]. There are many data mining algorithms for classification process such as: Naive Bayes, Rule Induction, K-NN, Random Forest, Support Vector Machine, Neural Network, and Decision Tree. In our research we used four of them: Naive Bayes, Rule Induction, K-NN, and Decision Tree.

Naive Bayes method depends basically on classifying a given document in any category using the joint probabilities of words and categories. The method called naive because of the assumption of word independence. It assumes that conditional probability of a word given a category should be independent from the conditional probabilities of the other words given the same category [12].

Rule Induction method represented the training data set using logical expressions as follows: IF (conditions) THEN (decision class). The left side of the expression (conditions) contains a set of attribute values that achieve the class attribute

978-1-5090-6538-7/17 $31.00 © 2017 IEEE
DOI 10.1109/PICICT.2017.24

value in the right side (decision class). The method classifies new rows depending on these expressions [13].

K-Nearest Neighbor (K-NN) method depending on learning by comparing given test row with training rows (training data set) which are similar to it. The training data set contains n attributes each record (row) represents point in n-dimensional space. This method firstly computes the distances between the new row and all the rows in the training data set. Then sort distances in increasing order and select the similar rows that have the smallest distances. Finally the classifier added the new row to the largest cluster form the k similar [12].

Decision Tree is a flowchart model somewhat similar to tree structure. Branches of the tree represents the results of the test and leafs retain a class label. Method used measures to select the attribute which best partitions the rows into distinct classes during tree construction. Tree pruning attempts to detect and delete tree branches that may reflect noise or outliers in the training data to improving classification accuracy [12].

## III. Related Work

Usability issue and recommender systems using data mining methods has been studied by several researchers. These researchers have proposed many approaches and techniques. And these some related researcher's works:

Ordonez et al in [5] studied categories that used from people to name objects, in particular images objects and learn models to predict categories for this images automatically. These models combine between visual recognition predictions and linguistic resources for words naturalness using the huge amount of text on the web.

Stocky et al in [6] proposed an approach to predict the text entered by users based on common sense reasoning. The proposed system can predict words based on their first few letters depending on a large-scale semantic network (OMCSNet). OMCSNet contains more than 280,000 commonsensical semantic relationships depending on around 700,000 English sentences.

Pachidi et al. in [7] proposed an approach for automating filling in forms depending on the patterns that appears from the analysis of data requested in forms to provide the exchange of data between user's Personal Information Management Systems and Web forms.

Al-Safadi et al. in [8] introduced an Arabic language application that addresses sentences autocompleting problem. This application provides fast suggestion retrieval time and also presents technique offer suggestions and ranking these suggestions.

Pachidi et al. in [9] presented method that depends on data mining techniques to analysis data that collected during software operation to extract knowledge about the usage of this software. The approach was evaluated through a prototype that was executed in the main online financial management application in the Netherlands (Exact Online).

Our work however, is different in the context that we try to design prediction model that can help us to make government decision system more user-friendly by suggesting entries to the system user during the entry process.

## IV. Methodology

We study some attributes from the database of Palestinian Government decisions system. Particularly we are trying to find a relationship between the placement sides and decision classification on one hand, and the relevant authorities on the other hand to find the patterns for suggesting entries to the user. Figure 1 show example of placement sides (on the right bottom side of the figure), decision classification (on the top right side of the figure) and the relevant authorities (on the top left side of the figure) in the system.

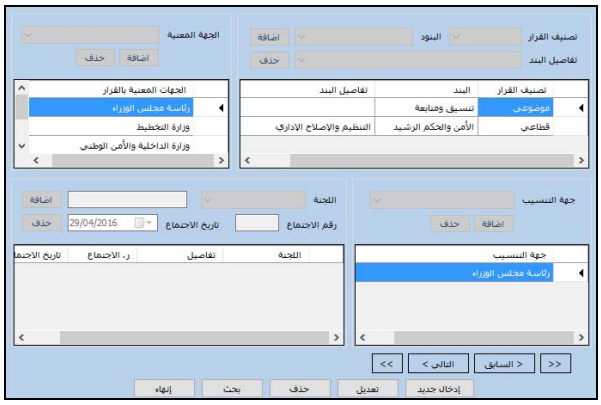


Figure 1. Example of placement sides (on the right bottom side of the figure), decision classification (on the top right side of the figure) and the relevant authorities (on the top left side of the figure) – Government Decision System.

### A. Data Set

The dataset used in our work was collected from Government decisions system of Palestinian Ministers Council in Gaza. The dataset consist of 13 attributes and contain 10238 rows. Table I show data set description.

TABLE I. Data Set Description

| Attribute | Details | The possible values |
|---|---|---|
| GDId | Government decisions ID in database, Primary key of table in database | Serial number |
| GovernmentId | Sequential number increases when a new government formed. The number is constant for all decisions of the same Government. | Numbers from 1 to 11 |
| SessionId | Sequential number increases with each weekly meeting of Government. The number is constant for all decisions of each meeting. | Serial number |
| DecisionId | Sequential number increases with each new decisions of weekly meeting of Government. | Serial number |
| DecisionDate | Date of Issuing each Government decisions. | Date values |
| Type | Authority that issued the decision the Prime Minister or Ministers Council. | two categorical values |
| PSide | Placement side. The side that requested issuance of the decision. | 37 categorical values |

| Attribute | Details | The possible values |
|---|---|---|
| RSide | Relevant side. The relevant authorities of the decision. | 37 categorical values |
| DClass | Main classification of the decision. | two categorical values |
| DCItem | Subclass of main classification of the decision. | 12 categorical values |
| DCIDetails | Class details of decision subclass. | 32 categorical values |
| Com | Government committee that have a role in the issuance of decision. | 5 categorical values |
| MeetingNum | Sequential number increases with each meeting of committee. | Serial number |

*B. Preparation and Preprocessing*

Data preparation and preprocessing is a necessary step for serious, effective and real-world data mining to increase the accuracy of the mining [14]. In our work we used three methods from RapaidMiner to prepare the data for mining process. The three methods are: Replace Missing Values, Select Attributes, and Filter Examples. These methods were used consecutively.

We applied Replace Missing Values method on data set to replace missing values in each attribute by the word " غير مصنف", which means unclassified. Then we noticed that group of old data were not classified. Figure 2 shows the result of applying "Replace Missing Values" method on the data set.

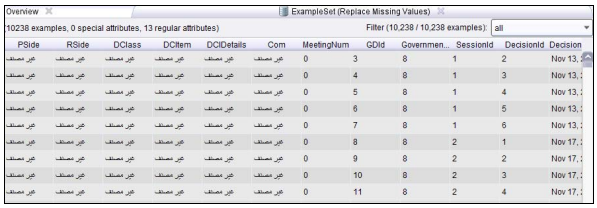
Figure 2. Result of applying Replace Missing Values method on the dataset.

Then we applied Select Attributes method on data set to select important attributes and ignored the rest. The selected attributes are: Type, DClass, DCItem, DCIDetails, PSide, RSide, and Com. Figure 3 shows partial results of applying "Select Attributes" method on the data set.

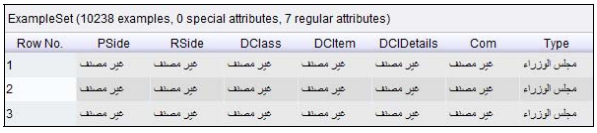
Figure 3. Result of applying Select Attributes method on the dataset.

Also we applied Filter Examples method on data set to exclude rows contain value "غير مصنف" in the five attributes (PSide, RSide, DClass, DCItem, DCIDetails) so as not to affect the accuracy of the results. Filter Examples method not applied to the two attributes DCIDetails and Com because the value "غير مصنف" represent 49.9% of DCIDetails attribute values and 73% of Com attribute values. In case of filtered the two attributes a large number of data rows lost. Figure 4 shows the result of applying the method on the data set.


Figure 4. Result of applying Filter Examples method on the dataset.

At the end of prepressing and preparation step we had data set consist of 7 attributes without missing values.

*C. Association Rules*

We used FP-Growth and Create Association Rules methods from RapaidMiner to create association rules in order to search for interesting relationship among dataset features in a proactive step before classification to ensure the existence of patterns that can depend on it in training the classifier.

Initially the output of preparation and preprocessing process are used as input of the creation association rules process. Then FP-Growth method is used to calculates all frequent itemsets from the dataset. Finally Create Association Rules method used to generates association rules from the frequent itemsets. After the completion of applying these methods we obtained set of rules as results. Firstly a general pattern appears through some of these rules. This pattern show that the placement side of each government decision is also be relevant authority of the same decision. For example if PSide = "سلطة الأراضي" then RSide = "سلطة الأراضي" cf (0.899) as shown in Figure 5.

| No. | Premises | Conclusion | Confidence |
|---|---|---|---|
| 31 | PSide = سلطة الأراضي | RSide = سلطة الأراضي | 0.899 |
| 30 | RSide = وزارة الشباب والرياضة | PSide = وزارة الشباب والرياضة | 0.871 |
| 29 | PSide = سلطة المياه | RSide = سلطة الأراضي | 0.851 |
| 28 | RSide = مجلس القضاء الأعلى | PSide = مجلس القضاء الأعلى | 0.700 |
| 27 | RSide = وزارة الزراعة | PSide = وزارة الزراعة | 0.676 |
| 26 | RSide = وزارة النقل والمواصلات | PSide = وزارة النقل والمواصلات | 0.662 |
| 25 | RSide = وزارة الأشغال العامة والإسكان | PSide = وزارة الأشغال العامة والإسكان | 0.647 |
| 24 | PSide = وزارة المالية | RSide = وزارة المالية | 0.634 |
| 23 | RSide = وزارة الاتصالات وتكنولوجيا المعلومات | PSide = وزارة الاتصالات وتكنولوجيا المعلومات | 0.634 |
| 22 | RSide = وزارة الاقتصاد الوطني | PSide = وزارة الاقتصاد الوطني | 0.631 |
| 21 | PSide = ديوان الموظفين العام | RSide = ديوان الموظفين العام | 0.624 |

Figure 5. Results of Association Rules when placement side of each government decision is also be relevant authority.

Also Figure 6, 7, and 8 shows some patterns that appeared in the results. One of these patterns was the relevant authority "سلطة الأراضي" appears frequently with placement sides (" سلطة الأراضي", "وزارة التربية والتعليم", "سلطة المياه", "وزارة الأوقاف والشؤون الدينية"). For example if PSide = "سلطة المياه" then RSide = " سلطة الأراضي" cf (0.851) as shown in Figure 6.

The second example show that relevant authority " وزارة المالية" appeared frequently with placement sides ("وزارة المالية", "وزارة النقل ", "وزارة الأشغال العامة والإسكان", "وزارة الشؤون الإجتماعية", "رئيس الوزراء", "وزارة الحكم المحلي", "والمواصلات"). For example if PSide = "وزارة الشؤون الإجتماعية" then RSide = "وزارة المالية" cf (0.331) as shown in Figure 7.

Also the relevant authority "ديوان الموظفين العام" appears frequently with placement sides ("ديوان الموظفين العام", " رئاسة الوزراء", "مجلس الوزراء", "وزارة الصحة", "وزارة التربية والتعليم"). For example

if PSide = "وزارة الصحة" then RSide = "ديوان الموظفين العام " cf (0.256) as shown in Figure 8.

| No. | Premises | Conclusion | Confidence |
|---|---|---|---|
| 75 | PSide = سلطة الأراضي | RSide = سلطة الأراضي | 0.899 |
| 73 | PSide = سلطة المياه | RSide = سلطة الأراضي | 0.851 |
| 33 | PSide = وزارة الأوقاف والشؤون الدينية | RSide = سلطة الأراضي | 0.325 |
| 25 | PSide = وزارة التربية والتعليم | RSide = سلطة الأراضي | 0.199 |

Figure 6. Result of Association Rules when relevant authority is " سلطة الأراضي".

| No. | Premises | Conclusion | Confidence |
|---|---|---|---|
| 68 | PSide = وزارة المالية | RSide = وزارة المالية | 0.634 |
| 34 | PSide = وزارة الشؤون الاجتماعية | RSide = وزارة المالية | 0.331 |
| 28 | PSide = وزارة الأشغال العامة والإسكان | RSide = وزارة المالية | 0.247 |
| 26 | PSide = وزارة النقل والمواصلات | RSide = وزارة المالية | 0.206 |
| 24 | PSide = وزارة الحكم المحلي | RSide = وزارة المالية | 0.199 |

Figure 7. Result of Association Rules when relevant authority is " وزارة المالية".

| No. | Premises | Conclusion | Confidence |
|---|---|---|---|
| 65 | PSide = ديوان الموظفين العام | RSide = ديوان الموظفين العام | 0.624 |
| 29 | PSide = وزارة الصحة | RSide = ديوان الموظفين العام | 0.256 |
| 27 | PSide = رئاسة مجلس الوزراء | RSide = ديوان الموظفين العام | 0.211 |
| 23 | PSide = وزارة التربية والتعليم | RSide = ديوان الموظفين العام | 0.192 |

Figure 8. Result of Association Rules when relevant authority is " ديوان الموظفين العام".

### *D. Classification*

Classification used to predict the relevant authorities of the decision "RSide" depending on the side that requested issuance of the decision "PSide" and the classification of the decision "DClass" "DCItem" "DCIDetails" so our class was "RSide".

Four classification methods from RapaidMiner are applied on the dataset: Naive Bayes, Rule Induction, K-NN, and Decision Tree.

Firstly output of preparation and preprocessing process are used as input to classification process. Then Set Role method is used to determine our labeled class. Finally Split Validation method used to splits up the Data set into a training set and test set and evaluates the model. This method performs a split validation in order to estimate the performance and used to estimate how accurately a model will perform in practice. Classification model using Naive Bayes, Rule Induction, K-NN, and Decision Tree was applied on the data.

## V. RESULTS AND DISCUSSION

Our work is classified into four main steps first of them data collection to create data set, the second step was preparation and preprocessing data set to increase the accuracy of the mining, the third step was creating association rules to search for interesting relationship among dataset features, and the last step was used classification methods (Naive Bayes, Rule Induction, K-NN, and Decision Tree) to achieve our goal.

After applying the classification techniques, we found variation in the accuracy and execution time results of the four classification techniques. The results were as follows:

Naive Bayes model gave the highest accuracy (68.41%) followed by K-NN (K value = 21) in the second place with accuracy (63.19%) followed by Decision Tree with accuracy (57.97%) and finally Rule Induction model gave the lowest accuracy (57.10%). Table II shows the classification models accuracy.

TABLE II. CLASSIFICATION MODELS ACCURACY

| Classification Model | Accuracy |
|---|---|
| Naive Bayes | 68.41% |
| Rule Induction | 57.10% |
| K-NN | 63.19% |
| Decision Tree | 57.97% |

Results in [15] also show the same disparity in accuracy rates among the four algorithms (see Table III) when used with different dataset.

TABLE III. CLASSIFICATION ALGORITHMS ACCURACY RESULTS IN [14]

| Algorithm | Naive Bayes | Rule Induction | K-NN | Decision Tree |
|---|---|---|---|---|
| Accuracy | 90.20% | 86.40% | 88.80% | 88.40% |

Accuracy rate in general may appear somewhat low, but justified the existence of a large range of values in class attribute not only true or false. Add to that each government decision has several relevant authorities not only one so it is natural that we do not find high accuracy percentage where the percentage distributed among the various relevant authorities.

Although the accuracy of the classification is not high, this accuracy rate may be relatively suitable for our user-entries prediction model. Because the system user essentially enters all the options by himself and our proposed model relieves the user of entry about 68.41% of the data, a satisfactory ratio.

## VI. CONCLUSION

In our work we have analyzed the stored data in the database of Palestinian Government decisions system and studied the relationship between some attributes. A prediction model has designed depending on the patterns that we have found in the stored data. This model may make the entry system more user-friendly in which it can suggest entries to the system user during data entry process.

Naive Bayes, Rule Induction, K-NN, and Decision Tree methods have been applied to the stored data. The experiment results have shown that Naïve Bayes is the best model among the other techniques by achieving the highest accuracy of 68.41% followed by K-NN in the second place with accuracy of 63.19% then Decision Tree with accuracy of 57.97% and finally Rule Induction model gave the lowest accuracy 57.10%.

In future we will apply this prediction model on the Government decisions system of Palestinian Ministers Council in Gaza for examination it in practice.

## REFERENCES


[1] Preece, Jenny, et al. Human-computer interaction. Addison-Wesley Longman Ltd., 1994.

[2] ISO 9241-11. Ergonomic requirements. Part 11: Guidance on usability, 1998

[3] ISO 13407. Human centered design processes for interactive systems, 1999.

[4] Data Mining: What is Data Mining, available on April 2016. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm.

[5] Ordonez, V., Liu, W., Deng, J., Choi, Y., Berg, A.C. and Berg, T.L., Predicting entry-level categories. International Journal of Computer Vision, 2015, 115(1), pp.29-43.

[6] Stocky, T., Faaborg, A. and Lieberman, H., April. A commonsense approach to predictive text entry. In CHI'04 Extended Abstracts on Human Factors in Computing Systems ACM, 2004, pp. 1163-1166.

[7] Winckler, M., Gaits, V., Vo, D.B., Sergio, F. and Rossi, G., An approach and tool support for assisting users to fill-in web forms with personal information. In Proceedings of the 29th ACM international conference on Design of communication, 2011, pp. 195-202.

[8] Al-Safadi, L., Aldayel, A., Alshaiban, A., Almasad, L. and Alromaih, S., A User Behavior Approach for phrase Autocompletion. Computers and Mathematics in Automation and Materials Science, 2014, pp. 108-112.

[9] Pachidi, S., Spruit, M. and Van De Weerd, I., Understanding users' behavior with software operation data mining. Computers in Human Behavior, 2014, 30, pp.583-594.

[10] Kantardzic, M., Data mining: concepts, models, methods, and algorithms. John Wiley & Sons, 2011.

[11] Pang-Ning, T., Steinbach, M. and Kumar, V., Introduction to data mining. In Library of congress, 2006, (Vol. 74).

[12] Entezari-Maleki, R., Rezaei, A. and Minaei-Bidgoli, B., Comparison of classification methods based on the type of attributes and sample size. Journal of Convergence Information Technology, 2009, 4(3), pp.94-102.

[13] Stefanowski, J. and Nowaczyk, S., An experimental study of using rule induction algorithm in combiner multiple classifier. International Journal of Computational Intelligence Research, 2007, 3(4), pp.335-342.

[14] Chandrasekar, P., & Qian, K., The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier. In Computer Software and Applications Conference (COMPSAC), 2016 IEEE 40th Annual Vol. 2, pp. 618-619.

[15] Goyal, V.K., A Comparative Study of Classification Methods in Data Mining using RapidMiner Studio. (IJIRSE) International Journal of Innovative Research in Science & Engineering, ISSN (Online), 2013, pp.2347-3207.