Applying Association Data Mining Technique in Civil Registration Data

Atif Ali Mohammed[1], Murtada El-mukashfi El-taher[2], Nora kamal AbdElgalil[3]

[1]Assistant Professor - Faculty of CS & IT - University Of Science & Technology - Khartoum – Sudan
[2]Assistant Professor - Faculty of Science - University Of Bakhalruda - Elduwiam – Sudan
[3]Civil Registry - Khartoum – Sudan

مستخلص :
تنقيب البيانات هو عملية إستقراء المعرفة من قواعد البيانات الضخمة ومستودعات البيانات والتي لم تكن معروفة مسبقا ، وان هذه المعلومات مفيدة للإستخدام في شتي المجالات (التنبوء |، دعم القرار ، تحليل الإتجاهات) ، تعتبر عمليات تقيب البيانات من الأداوت والمناهج القوية جدا في إكتشاف العلاقات بين البيانات المخزنة في مستودعات وقواعد البيانات الضخمة . هدفت هذه الدراسة إلى تطبيق أحد تقنيات تنقيب البيانات (Association) في قواعد بيانات السجل المدني السوداني لغرض إستخراج معلومات من هذه البيانات الضخمة التي يمكن أن تفيد صانعي القرار السودانيين في مجالات شتى خاصة في ما يتعلق بالأمن القومي من تعليم وصحة وغيرها. تم جمع البيانات من بعض مراكز السجل المدني وتم إعدادها وطبقت عليها خوارزمية (Apriori)، بإستخدام الحزمة (WEKA) وتم إستخراج علاقات بين بيانات التعليم والعمل . النتائج المستخلصة عبرت عن نسب التعليم والعمالة في الولايات التي طبقت عليها الخوارزمية ،والتي يمكن أن يستفاد منها في دعم قرارات أصحاب الشأن في هذا المجال ، كما أن النتائج التي تم الحصول عليها تثبت مدى جدوى تطبيق تقنيات التنقيب في مثل هذا النوع من المجالات.
كلمات مفتاحية : تنقيب البيانات، جمعية، السجل المدني، خوارزمية Apriori، WEKA

Abstract:
Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. Nowadays data mining is a modern and powerful tool, automat zing the process of discovering relationships and combinations in raw data and using the results in an automatic decision support. This study is focus on finding relationships and patterns among Sudanese citizens data, this relations can be used to enhance and develop the whole states of Sudan in educational field and also providing best jobs opportunities that can be useful for both Sudanese citizens and government by providing statistical information to decision makers on status of education and jobs. WEKA Machine learning tool used in our experiments, then applying simple Apriori algorithm in Civil Registration Dataset which it's represent different Sudanese citizens depending on their state, after using this algorithm and WEKA tool extract a lot of relations in Educational field and works status of specific state and describe this data in statistical view to help decision makers to take the optimal solutions for problems of different Sudanese states.

Keywords: Data mining**,** Association, Civil Registration, Apriori algorithm, WEKA.

1. Introduction

Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large scale data. Association rule mining finds interesting association or correlation relationships among a large set of data items. With massive amounts of data continuously being collected and stored in databases, many industries are becoming interested in mining association rules from their databases. This study focuses on using data mining techniques (Association) to study and analyze the huge data of Sudanese nation and then generate rules that can tell us which State needs to be developed in specific field such as education, criminal and jobless problems just by using this model and then facing this issue.

2. Problem statement

Sudanese Civil Registration Centers stores huge data inside their databases, this data include important information about Sudanese citizens, which can be use as strategic

information in different issues such as education, health or any other national security issues. Using traditional tools to extract information from this huge data repositories faces many difficulties and challenges, so the main problem this paper discuss is how to make use of this huge data using advance analysis tools. and try to get maxim benefit by used this huge data.

3. Objective of the study.

The main objective of this study is to apply association data mining technique in Sudanese civil registration data repositories to find relations and rules among this data to be use as information can help decision makers in some national security issues. For this purpose, the study uses selected attributes from the hole attributes that Sudanese civil registration database contains such as education, job and gender, and selected states for applying the techniques, then uses the model as a sample of study to be apply in other Sudan states. As a result of applying the output of this study, there will be chance to enhance education strategies and Jobs Opportunities in all states.

4. Methodology and Tool;

As a methodology, CRISP-DM (Cross Industry Standard Process for Data Mining) steps followed. Then WEKA Machine learning tool was used to apply the association rule algorithm. WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

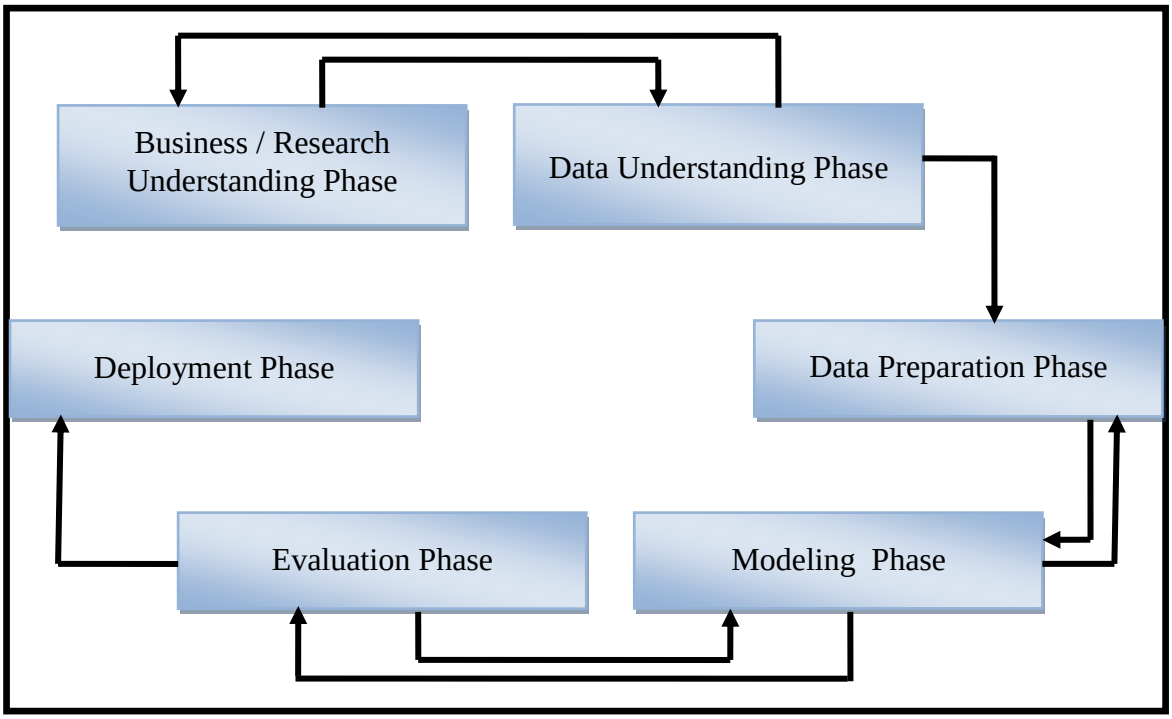

Figure (1) CRISP steps

5. Main Idea :

this study is trying to study Sudanese data to develop a model that can be used by decision makers and determine the number of rules and choose one rule that is the best to make Sudanese different states in the good categories of Education and jobs

opportunities, by using Association which it's data technique to study this huge data, that can't be study using traditional tools.

6. Literature Review :

This section summarizes various review and technical articles on data mining techniques. Several works have been carried out by many researchers. This section presents a brief summary on the basis of literature.

Moawia and other tried to put a new direction for the evaluation of some techniques for solving data mining tasks. The new approach has succeed in defining some new criteria for the evaluation process, and it has obtained valuable results based on what the technique is, the environment of using each techniques, the advantages and disadvantages of each technique, the consequences of choosing any of these techniques to extract hidden predictive information from large databases, and the methods of implementation of each technique. [Moawia 2010]

Kavitha Sasipraba evaluated the performance of distributed data mining framework on Java platform. Association rule mining was used for discovering interesting patterns from a large amount of data. [Kavitha 2011]

Manoj Bala applied an application of data mining in educational institute to extract the useful information from the huge dataset and provided analytical tool to view and used this information for decision making process. They also conducted a research on student learning result based on data mining.[Manoj 2012]

7. Review of Data Mining

7-1. Data Mining :

Data mining refers to extracting useful information from vast amounts of data. Many other terms are being used to interpret data mining, such as knowledge mining from databases, knowledge extraction, data analysis, and data archaeology. Nowadays, it is commonly agreed that data mining is an essential step in the process of knowledge discovery in databases. [Yihao 2010]

Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data. [Srinivasa 2015]

Another Definition of Data Mining is about processing data and identifying patterns and trends in that information so that you can decide or judge. Data mining principles have been around for many years, but, with the advent of big data, it is even more prevalent. [Martin 2012]

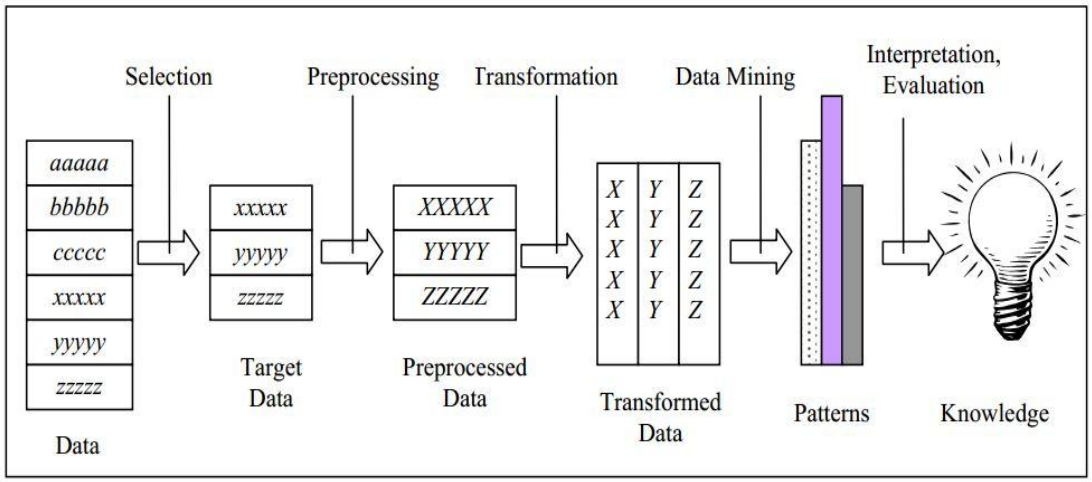

Figure (2) : Data Mining Process

The first three steps in Figure (2) involve preparing the data for mining. The relevant data must be selected from a potentially large and diverse set of data, any necessary preprocessing must then be performed, and finally the data must be transformed into a representation suitable for the data mining algorithm that is applied in the data mining step. [Gary 2010]

I can say that Data mining is a computerized technology that uses complicated algorithms to find relationships in large data bases Extensive growth of data gives the motivation to find meaning full patterns among the huge data set.

7-2. Data Mining Tools: Data mining is not all about the tools or database applications or software that are being used we can perform data mining with comparatively modest database systems and simple tools, including creating and writing our own, or using existence software packages.

7-3. Data Mining techniques: Several core techniques that are used in data mining describe the type of mining and data recovery operation. Unfortunately, the different companies and solutions do not always share terms, which can add to the confusion and apparent complexity. [Martin 2012]

7-3-1. Association : or relation is probably the better known and most familiar and straightforward data mining technique. here you make a simple correlation between two or more items, often of the same type to identify patterns.

7-3-2. Classification: you can use classification to build up an idea of the type object by describing multiple attributes to identify a class. Additionally, you can use classification as a feeder to, or the result of, other techniques. For example, you can use decision trees to determine a classification. Clustering allows you to use common attributes in different classifications to identify clusters.

7-3-3. Clustering : by examining one or more attributes or classes, you can group individual pieces of data together to form a structure opinion. At a simple level, clustering is using one or more attributes as your basis for identifying a cluster of correlating results. Clustering is useful to identify different information because it correlates with other examples so you can see where the similarities and ranges agree.

7-3-4. Prediction : prediction is a wide topic and runs from predicting the failure of components or machinery, to identifying fraud and even the prediction of company profits. Used in combination with the other data mining techniques, prediction involves analyzing trends, classification, pattern matching, and relation. by analyzing past events or instances, you can make a prediction about an event

7-4. Data Matrix

Data can often be represented or abstracted as an n × d *data matrix*, with n rows and d columns, where rows correspond to entities in the dataset, and columns represent attributes or properties of interest. Each row in the data matrix records the observed attribute values for a given entity. The n× d data matrix is given as [Mohammed 2014]

$$\mathbf{D} = \begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}$$

Figure (2) : Data Matrix

7-5. Data Preparation

Some data preparation is needed for all mining tools, the purpose of preparation is to transform data sets so that their information content is best exposed to the mining tool, Error prediction rate should be lower (or the same) after the preparation as before it. Preparing data also prepares the miner so that when using prepared data, the miner produces better models, faster. [Jiawei 2012]

7-6. Data Mining Applications and usage :

The data mining applications can be generic or domain specific. The generic application is required to be an intelligent system that by its own can takes certain decisions like: selection of data, selection of data mining method, presentation, and interpretation of the result. Some generic data mining applications cannot take its own these decisions but guide users for selection of data, selection of data mining method and for the interpretation of the results. [Rupali 2013]

7-7. Data Mining Algorithms :

In general terms, data mining comprises techniques and algorithms, for determining interesting patterns from large datasets. There are currently hundreds of algorithms that perform tasks such as frequent pattern mining, clustering, and classification, among others. Understanding how these algorithms work and how to use them effectively is a continuous challenge faced by data mining analysts, researchers, and practitioners, because the algorithm behavior and patterns it provides may change significantly as a function of its parameters.

7-7-1. C4.5 Algorithm : C4.5 constructs a classifier in the form of a decision tree. In order to do this, C4.5 is given a set of data representing things that are already classified. Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs. [Bharti 2014]

7-7-2. SVM Algorithm : Support vector machine (SVM) learns a hyper plane to classify data into 2 classes. At a high-level, SVM performs a similar task like C4.5 except SVM doesn't use decision trees at all. In today's machine learning applications, support vector machines are considered a must try—it offers one of the most robust and accurate methods among all well-known algorithms. It has a sound theoretical foundation, requires only a dozen examples for training, and is insensitive to the number of dimensions. In addition, efficient methods for training SVM are also being developed at a fast pace**.** [Madhes 2016]

7-7-3. Apriori Algorithm : One of the most popular data mining approaches is to find frequent itemsets from a transaction dataset and derive association rules. Finding frequent itemsets (item sets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent item sets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Apriori is a seminal algorithm for finding frequent itemsets using candidate generation. The Apriori algorithm learns association rules and is applied to a database containing a large number of transactions. [Madhes 2016]

7-7-4. k-means Algorithm : is a simple iterative method to partition a given dataset into a user specified number of clusters, k. k-means creates k groups from a set of objects so that the members of a group are more similar. It's a popular cluster analysis technique for exploring a dataset. [Madhes 2016]

7-7-5. EM Algorithm : Expectation-Maximization (EM) is generally used as a clustering algorithm (like k-means) for knowledge discovery. In statistics, the EM algorithm iterates and optimizes the likelihood of seeing observed data while estimating the parameters of a statistical model with unobserved variable. [Madhes 2016]

7-7-6. Naive Bayes Algorithm : is not a single algorithm, but a family of classification algorithms that share one common assumption: Every feature of the data being classified is independent of all other features given the class. Two features are independent when the value of one feature has no effect on the value of another feature. [Madhes 2016]

7-7-7. PageRank Algorithm : is a search ranking algorithm using hyperlinks on the Web. Page Rank produces a static ranking of Web pages in the sense that a Page Rank value is computed for each page off-line and it does not depend on search queries. It's a type of network analysis looking to explore the associations among objects. [Phyu 2013]

7-7-8. kNN Algorithm : k-Nearest Neighbors, is a classification algorithm. However, it differs from the AdaBoost classifiers described because it's a lazy learner. A lazy learner doesn't do much during the training process other than store the training data. Only when new unlabeled data is input does this type of learner look to classify. [Madhes 2016]

7-7-9. CART Algorithm : stands for classification and regression trees. It is a decision tree learning technique that outputs either classification or regression trees. Like C4.5, CART is a classifier. A classification tree is a type of decision tree. The output of a classification tree is a class. [Bhumika 2017]

7-8. Data Mining Software :

There are many software that can be used to apply data mining algorithms including:

7-8-1. WEKA : is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. WEKA is open source software issued under the GNU General Public License for data mining. [Weka 2017]

7-8-2. KNIME : an open source data integration, processing, analysis, and exploration Analytics Platform to leading open solution for data-driven innovation, designed for discovering the potential hidden in data mining for fresh insights, or predicting new futures. [KNIME 2017]

7-8-3. Mahout : The Mahout machine learning library is mining large data sets. It supports recommendation mining, clustering, classification and frequent itemset mining. and build an environment for quickly creating scalable perform machine learning applications. [Madhes 2016]

7-8-4. Rattle : is a popular GUI for data mining using R. It presents statistical and visual summaries of data, transforms data so that it can be readily modeled, builds both unsupervised and supervised machine learning models from the data, presents the performance of models graphically, and scores new datasets for deployment into production. [Rattle 2017]

7-8-5. R-Programming : The R Project for Statistical Computing. [Madhes 2016]

7-8-6. Orange : an open source machine learning and data visualization for novice and expert. Interactive data analysis workflows with a large toolbox. [Orange 2017]

7-8-7. Scikit-learn : an open source Machine Learning in Python**.** and Simple and efficient tools for data mining and data analysis. [Scikit-learn 2017]

7-8-8. MATLAB : Analyze and design the systems and products transforming our world. [Madhes 2016]

7-8-9. LIBSVM : A Library for Support Vector Machines. is an integrated software for support vector classification, regression and distribution estimation. It supports multi-class classification. [LIBSVM 2017]

Table 1. Show Data Mining Algorithm and Software Used

Data Mining Algorithm	Algorithm used for	Software Used
C4.5 Algorithm	classification	Orange
SVM Algorithm	classification or regression	MATLAB, LIBSVM tools
Apriori Algorithm	Boolean Association	Weka, and Orange tools
k-means Algorithm	Clustering	Weka, MATLAB
EM Algorithm	classification or regression	Weka, modules in R & scikit-learn tools.
Naive Bayes Algorithm	classification	Orange, scikit-learn, Weka and R
PageRank Algorithm	classification	The network analysis package R
kNN Algorithm	Classification	MATLAB, scikit-learn,
CART Algorithm	classification or regression	scikit-learn, R package, Weka and MATLAB

There are many tools & algorithms are there for data mining and I have collected at very high level for top tools & algorithms which are currently used in market for data mining.

8. Second: Implementation

8-1. Association Rule Mining Algorithms :

An association rule is a rule which implies certain association relationships among a set of objects (such as "occur together" or "one implies the other") in a database. Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X Y, where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y. [Rashmi 2014]

An example of an association rule is:"30% of farmers that grow wheat also grow pulses; 2% of all farmers grow both of these items". Here 30% is called the confidence of the rule, and 2% the support of the rule. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints.

Association Rule An association rule is an implication expression of the form X $\rightarrow$Y, where X and Y are disjoint itemsets, i.e., $X \cap Y = \emptyset$. The strength of an association rule can be measured in terms of its support and confidence. Support determines how often a rule is applicable to a given data set, while confidence determines how frequently items in Y appear in transactions that contain X. The formal definitions of these metrics are : [Kumar 2005]

$$\text{Support, } s(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

$$\text{Confidence, } c(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

8-2. Support and Confidence :

Support : is an important measure because a rule that has very low support may occur simply by chance. A low support rule is also likely to be uninteresting from a business perspective because it may not be profitable to promote items that customers seldom buy together. For these reasons, support is often used to eliminate uninteresting rules. support also has a desirable property that can be exploited for the efficient discovery of association rules.

Confidence : on the other hand, measures the reliability of the inference made by a rule. For a given rule X −→ Y, the higher the confidence, the more likely it is for Y to be present in transactions that contain X. Confidence also provides an estimate of the conditional probability of Y given X. Association analysis results should be interpreted with caution. The inference made by an association rule does not necessarily imply causality. Instead, it suggests a strong co-occurrence relationship between items in the antecedent and consequent of the rule. Causality, on the other hand, requires knowledge about the causal and effect attributes in the data and typically involves relationships occurring over time. [Kumar 2005]

8-3. Formulation of Association Rule Mining Problem :

The association rule mining problem can be formally stated as follows as given a set of transactions T, find all the rules having support ≥ min-sup and confidence ≥ min-conf ,where min-sup and min-conf are the corresponding support and confidence thresholds.

A brute-force approach for mining association rules is to compute the support and confidence for every possible rule. This approach is prohibitively expensive because there are exponentially many rules that can be extracted from a data set. More specifically, the total number of possible rules extracted from a data set that contains d items is : [Kumar 2005]

$$\mathbf{R = 3^d - 2^{d+1} + 1}$$

8-4. Apriori Algorithm :

Aprioris a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules [AS94b]. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties, as we shall see later. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore. kC 1/-itemsets. First, the set of frequent1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted by L_1.Next, L_1 is used to find L_2, the set of frequent 2-itemsets, which is used to find L_3, and so on, until no more frequent k-itemsets can be found. The finding of each L_k requires one full scan of the database.
To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property is used to reduce the search space. [Jiawei 2012]

8-5. Apriori property :

All nonempty subsets of a frequent itemset must also be frequent. The Apriori property is based on the following observation. By definition, if an itemset *I* does not satisfy the minimum support threshold, min-sup, then *I* is not frequent, that is, P(*I*) < min-sup. If an item A is added to the itemset *I*, then the resulting itemset ($I \cup A$) cannot occur more frequently than *I*. Therefore, $I \cup A$ is not frequent either that is, P($I \cup A$) < min-sup. This property belongs to a special category of properties called **antimonotone** in the sense that if a set cannot pass a test, all of its supersets will fail the same test as well. It is called antimonotone because the property is monotonic in the context of failing a test., let us look at how L_{k-1} is used to find L_k for k ≥ 2. A two-step process is followed, consisting of **join** and **prune** actions. [Kumar 2005]

The join step : To find L_k, a set of **candidate** k-itemsets is generated by joining L_{k-1} with itself. This set of candidates is denoted C_k. Let L_1 and L_2 be itemsets in L_{k-1}. The notation $L_i[j]$ refers to the jth item in L_i and $L_1[k - 2]$ refers to the second to the last item in L_1. For efficient implementation, Apriori assumes that items within a transaction or

itemset are sorted in lexicographic order. For the (k- 1)-itemset, L_i, this means that the items are sorted such that $L_i[1] < L_i[2] < \cdots < L_i[k - 1]$. The join, $L_{k-1} \bowtie L_{k-1}$, is performed, where members of L_{k-1} are joinable if their first (k- 2) items are in common. That is, members L_1 and L_2 of L_{k-1} are joined if $(L_1[1] = L_2[1]) \wedge (L_1[2] = L_2[2]) \wedge \ldots \wedge (L_1[k-2] = L_2[k-2]) \wedge (L_1[k-1] < L_2[k-1])$. The condition $(L_1[k-1] < L_2[k-1]$ simply ensures that no duplicates are generated. The resulting itemset formed by joining L_1 and L_2 is $L_1[1], L_1[2], \ldots, L_1[k - 2], L_1[k - 1], L_1[k - 1]$.

The prune step : C_k is a superset of L_k, that is, its members may or may not be frequent, but all of the frequent k-itemsets are included in C_k. A scan of the database to determine the count of each candidate in C_k would result in the determination of L_k. C_k however, can be huge, and so this could involve heavy computation. To reduce the size of C_k, the Apriori property is used as follows. Any (k – 1)-itemset that is not frequent cannot be a subset of a frequent k-itemset. Hence, if any (k – 1)-subset of a candidate k-itemset is not in L_{k-1}, then the candidate cannot be frequent either and so can be removed from C_k. This subset testing can be done quickly by maintaining a hash tree of all frequent itemsets.

8-6. Datasets Description :

The dataset was been got from civil Registration unit, for educational purpose, the dataset is a combination of various tables in civil registration database they are four tables is (main Information table, Education table, Jobs table and state table). and bullied the structure of those table with (No, Name, Data type and constraint). then combine those tables together and make the lookup table from there attributes (gender, job and education). after that used this attribute to generate the model as described. And number of records provided by civil registration as follow:

Table 2. Show the records numbers table from civil registration

No	State	Number of Records
1	Gezira	4408
2	North Darfur	480
3	River Nile	3355
4	North Kordofan	596

8-7. Implementation of Apriori Algorithm on Dataset :

When Apriori algorithm run over the dataset, it generate a model contain the following information (Run Information, Apriori and Best rule found by selecting Apriori Algorithm).

Run information :

Relation

State	Gezira	North Darfur	River Nile	North Kordofan
Instances	4408	3355	480	596

Attributes: 3

Gender
Job
Education

Apriori :

State	Gezira	North Darfur	River Nile	North Kordofan

Minimum support:	0.2	0.2	0.25	0.25
Minimum metric <confidence>	0.4	0.4	0.4	0.4
Number of cycles performed	16	16	15	15

Generated sets of large item sets

State	Gezira	North Darfur	River Nile	North Kordofan
Size of set of large item sets L1:	7	7	6	6
Size of set of large item sets L2	7	7	7	5
Size of set of large item sets L3	1	1	-	-

9. Best rules found :

9-1. Gezira State :

- Gender=female Education=Uneducated 1001 ==> Job=Not Works 995 <conf:(0.99)> lift:(1.68) lev:(0.09) [401] conv:(58.17)
- Education=School 1059 ==> Job=Not Works 925 <conf:(0.87)> lift:(1.47) lev:(0.07) [296] conv:(3.19)
- Education=High Education 1114 ==> Job=Works 946 <conf:(0.85)> lift:(2.09) lev:(0.11) [492] conv:(3.91)
- Education=Uneducated 1807 ==> Job=Not Works 1439 <conf:(0.8)> lift:(1.34) lev:(0.08) [367] conv:(1.99)
- Gender=female 2252 ==> Job=Not Works 1637 <conf:(0.73)> lift:(1.23) lev:(0.07) [301] conv:(1.49)
- Job=Not Works Education=Uneducated 1439 ==> Gender=female 995 <conf:(0.69)> lift:(1.35) lev:(0.06) [259] conv:(1.58)
- Job=Works 1793 ==> Gender=male 1178 <conf:(0.66)> lift:(1.34) lev:(0.07) [301] conv:(1.49)
- Job=Not Works 2615 ==> Gender=female 1637 <conf:(0.63)> lift:(1.23) lev:(0.07) [301] conv:(1.31)
- Gender=female Job=Not Works 1637 ==> Education=Uneducated 995 <conf:(0.61)> lift:(1.48) lev:(0.07) [323] conv:(1.5)
-. Education=Uneducated 1807 ==> Gender=female 1001 <conf:(0.55)> lift:(1.08) lev:(0.02) [77] conv:(1.1)

9-2. River Nile State :

- Education=Baby Care 975 ==> Job=Not Works 975 <conf:(1)> lift:(1.57) lev:(0.11) [354] conv:(354.25)
- GENDER=female Education=Uneducated 765 ==> Job=Not Works 760 <conf:(0.99)> lift:(1.56) lev:(0.08) [272] conv:(46.33)
- Education=School 836 ==> Job=Works 829 <conf:(0.99)> lift:(2.73) lev:(0.16) [525] conv:(66.53)
- Education=Uneducated 1224 ==> Job=Not Works 1046 <conf:(0.85)> lift:(1.34) lev:(0.08) [266] conv:(2.48)
- Job=Not Works Education=Uneducated 1046 ==> GENDER=female 760 <conf:(0.73)> lift:(1.36) lev:(0.06) [199] conv:(1.69)
- GENDER=female 1799 ==> Job=Not Works 1303 <conf:(0.72)> lift:(1.14) lev:(0.05) [157] conv:(1.32)

- Job=Works 1219 ==> Education=School 829 <conf:(0.68)> lift:(2.73) lev:(0.16) [525] conv:(2.34)
- Education=Uneducated 1224 ==> GENDER=female 765 <conf:(0.63)> lift:(1.17) lev:(0.03) [108] conv:(1.23)
- Education=Uneducated 1224 ==> GENDER=female Job=Not Works 760 <conf:(0.62)> lift:(1.6) lev:(0.08) [284] conv:(1.61)
- Job=Not Works 2136 ==> GENDER=female 1303 <conf:(0.61)> lift:(1.14) lev:(0.05) [157] conv:(1.19)

9-3. North Darfur State :

- Job=Works 173 ==> Gender=male 163 <conf:(0.94)> lift:(1.41) lev:(0.1) [47] conv:(5.21)
- Gender=female 159 ==> Job=Not Works 149 <conf:(0.94)> lift:(1.47) lev:(0.1) [47] conv:(5.21)
- Education=High Education 195 ==> Gender=male 142 <conf:(0.73)> lift:(1.09) lev:(0.02) [11] conv:(1.2)
- Education=High Education 195 ==> Job=Not Works 134 <conf:(0.69)> lift:(1.07) lev:(0.02) [9] conv:(1.13)
- Education=Uneducated 212 ==> Job=Not Works 131 <conf:(0.62)> lift:(0.97) lev:(0.01-) [-4] conv:(0.93)
- Education=Uneducated 212 ==> Gender=male 128 <conf:(0.6)> lift:(0.9) lev:(0.03-) [-13] conv:(0.83)
- Job=Not Works 307 ==> Gender=male 158 <conf:(0.51)> lift:(0.77) lev:(0.1-) [-47] conv:(0.68)
- Gender=male 321 ==> Job=Works 163 <conf:(0.51)> lift:(1.41) lev:(0.1) [47] conv:(1.29)
- Gender=male 321 ==> Job=Not Works 158 <conf:(0.49)> lift:(0.77) lev:(0.1-) [-47] conv:(0.71)
- Job=Not Works 307 ==> Gender=female 149 <conf:(0.49)> lift:(1.47) lev:(0.1) [47] conv:(1.29)

9-4. North Kordofan State :

- job=Works 285 ==> GENDER=male 243 <conf:(0.85)> lift:(1.45) lev:(0.13) [75] conv:(2.74)
- GENDER=female 246 ==> job=Not Works 204 <conf:(0.83)> lift:(1.59) lev:(0.13) [75] conv:(2.74)
- EDUCATION=High Education 312 ==> GENDER=male 218 <conf:(0.7)> lift:(1.19) lev:(0.06) [34] conv:(1.36)
- GENDER=male 350 ==> job=Works 243 <conf:(0.69)> lift:(1.45) lev:(0.13) [75] conv:(1.69)
- job=Not Works 311 ==> GENDER=female 204 <conf:(0.66)> lift:(1.59) lev:(0.13) [75] conv:(1.69)
- GENDER=male 350 ==> EDUCATION=High Education 218 <conf:(0.62)> lift:(1.19) lev:(0.06) [34] conv:(1.25)
- job=Works 285 ==> EDUCATION=High Education 218 <conf:(0.55)> lift:(1.06) lev:(0.01) [8] conv:(1.06)
- EDUCATION=High Education 312 ==> job=Works 218 <conf:(0.51)> lift:(1.06) lev:(0.01) [8] conv:(1.05)
- job=Not Works 311 ==> EDUCATION=High Education 154 <conf:(0.5)> lift:(0.95) lev:(0.01-) [-8] conv:(0.94)

- EDUCATION=High Education 312 ==> job=Not Works 154 <conf:(0.49)> lift:(0.95) lev:(0.01-) [-8] conv:(0.94)

10. The results for Apriori algorithm are the following

The program generated the sets of large item sets found for each support size considered. In this case eight hundred eighty two item sets of two items were found to have the required Minimum support. By default Apriori tries to generate ten rules . It begins with a minimum support of 100% of the data items and decrease this in steps of 4% until there are at least ten rules with the required minimum confidence , or until the support has reached a lower bound of 10%whichever occur first . the minimum confidences is set 0.4 (40%). the minimum support decreased to 0.2 (20%) in the Gezira State and River Nile State. but 0.25 (25%) in the North Darfur State and North Kordofan State before the required number of rules can be generated . generation of the required number of rules involved a total of 16 iteration. The last part gives the association rules that are found . the number preceding ==> symbol indicates the rules support , that is the number of items covered by its premise . following the rule is the number of those items for which the rules consequent holds as well, In the parentheses there is a confidences of the rule.

11. Conclusion:

In this study, used data mining techniques Association Rule to generate model that can be used by decision makers to solve specific problem.

This study is seeking to provide to idea of what is going on in Sudan in many fields the Educational filed , working filed and Gendering field to let decision makers do the right thing for the place, by providing models that shows decision makers where is the problems and how they can have solved it by making the right decision depending on previous model

Table 3. Conclusion of Applying Association Data Mining Technique in Civil Registration Data

State	Education, Job	Education, Gender	Job, Gender
Gezira	School 1059 Not Works 925 High 1114 Works 946 Uneducated 1807 Not Works 1439	Uneducated 1807 female 1001 female 1637 Uneducated 995	Works 1793 male 1178 Not Works 2615 female 1637 female 2252 Not Works 1637
River Nile	School 836 Works 829 Baby Care 975 Not Works 975 Uneducated 1224, Not work 1046 Works 1219 School 829	Uneducated 1224, female 765	Not Works 2136 female 1303 female 1799 Works 1303
North Darfur	High Education 195 Not Works 134 Uneducated 212 Not Works 131	High Education 195 male 142 Uneducated 212 male 128	Works 173 male 163 female 159 Not Works 149 Not Works 307 male 158 male 321 Works 163
North Kordofan	Works 285 High Education 218 High Education 312 Works 218 Not Works 311 High Education 154	High Education 312 male 218 male 350 High Education 218	Works 285 male 243 female 246 Not Works 204 male 350 Works 243

12. Recommendations and future work.

Lastly I recommend the following:

Also Apriori algorithm as a data mining technique can help in finding strong relation among huge data with high accuracy and confidence, we recommend to use this algorithm in the hole civil data to help in designing strategies in many fields like health care, education, crime ….etc.

For the future we recommend applying hash base technique as data mining association another approach in the same data for the purpose of comparing association techniques, and chose which is better based on some important factors like accuracy.

References :

1. [Bharti 2014] Bharti Thakur & Manish Mann, Data Mining With Big Data Using C4.5 and Bayesian Classifier, ijarcsse Volume 4, Issue 8, 2014.
2. [Bhumika 2017] Bhumika Gupta, & others, Analysis of Various Decision Tree Algorithms for Classification in Data Mining, IJCA, Volume 163 – No 8, 2017, On-line available to https://pdfs.semanticscholar.org/fd39/e1fa85e5b3fd2b0d000230f6f8bc9dc694ae.pdf, up to date August 2017.
3. [Gary 2010] Gary M. Weiss, & Brian D. Davison, Data Mining To appear in the Handbook of Technology Management, H. Bidgoli (Ed.), John Wiley and Sons, 2010.
4. [Jiawei 2012] Jiawei Han, Micheline Kamber & Jian Pe, Data Mining Concepts and Techniques - Third Edition, Morgan Kaufmann Publishers is an imprint of Elsevier, 2012
5. [Kumar 2005] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Addison-Wesley, 2005. ISBN : 0321321367
6. [Kavitha 2011] Kavitha P., T. Sasipraba, "Performance Evaluation of Algorithms using a Distributed Data Mining Framework based on Association Rule Mining", International Journal on Computer Science & Engineering (IJCSE), 2011.
7. [KNIME 2017] KNIME Official Website, On-line available to https://www.knime.com/, 2017, up to date August 2017.
8. [LIBSVM 2017] LIBSVM Official Website, On-line available to https://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2017, up to date August 2017.
9. [Moawia 2010] Moawia Elfaki Yahia, A New Approach for Evaluation of Data Mining Techniques, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010.
10. [Manoj 2012] Manoj Bala, Study of Application of Data Mining Technique in Education", International Journal of Research in Science and Technology, Vol. No. 1, Issue No. IV, Jan-March, 2012.
11. [Martin 2012] Martin Brown- Data mining techniques 2012 On-line available to https://www.ibm.com/developerworks/library/ba-data-mining-techniques/ up to date August 2017.
12. [Mohammed 2014] Mohammed J. Zaki, Wagner Meira, Jr, Wagner Meira Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, 2014.
13. [Madhes 2016] Madhes DWBI, Data Mining Tools & Algorithms,2016 On-line available to https://dwbicastle.com/2016/12/15/data-mining-tools-algorithms/ up to date August 2017.
14. [Orange 2017] Orange Official Website, On-line available to https://orange.biolab.si/, 2017, up to date August 2017
15. [Phyu 2013] Phyu Thwe, Proposed Approach For Web Page Access Prediction Using Popularity And Similarity Based Page Rank Algorithm, IJSTR VOLUME 2, ISSUE 3, MARCH 2013, On-line available to http://www.ijstr.org/final-print/mar2013/, up to date August 2017

16. [Rupali 2013] Rupali Gaurav Gupta Data Mining: Techniques, Applications and Issues, IJARCSEE Volume 2, Issue 2, 2013.

17. [Rashmi 2014] Rashmi Jha & Amit Kumar, Association Rule Mining Algorithm In Parallel And Distributed Data Mining, International Multidisciplinary e-Journal, ISSN 2277 – 4262, 2014, On-line available to http://www.shreeprakashan.com/Documents/20140707132413535.1.Rashmi%20Jha..pdf, up to date August 2017.

18. [Rattle 2017] Rattle Official Website, On-line available to https://rattle.togaware.com/, 2017, up to date August 2017.

19. [Srinivasa 2015] K.G. Srinivasa, Anil Kumar Muppall Guide to High Performance Distributed Computing: Case Studies with Hadoop, Scalding and Spark, Springer, 2015

20. [Scikit-learn 2017] Scikit-learn Official Website, On-line available to http://scikit-learn.org/stable/, 2017, up to date August 2017

21. [Witten 1999] Witten, Ian and Eibe Frank, Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 1999

22. [Weka 2017] Weka Official Website, On-line available to http://www.cs.waikato.ac.nz/ml/weka/, 2017, up to date August 2017

23. [Yihao 2010] Yihao li , Data Mining: Concepts, Background, and Methods of Integrating, Uncertainty in Data Mining - On-line available to http://www.ccsc.org/southcentral/E-Journal/2010 up to date August 2017.