APPLICATION OF LABELLED K MEANS CLUSTERING FOR GIS CONTRACT AUTOMATION

Muhammad Shaheen

K Means Clustering is an unsupervised classification technique which is suitable for the dataset which do not have any class labels. The absence of labels constrained its application on different text data sets. Labelled K Means clustering generate labels for the clusters obtained from K Means Clustering which makes this technique more decisive for text data sets. A novel application of Labeled K Means Clustering for automation of technical part of GIS contracts is given in this paper. Geographic Information System (GIS) contract which is signed between GIS service provider and a client requires an efficient tracking system during all its vital phases (development, operation and maintenance). At present, the tracking is done in a non-discrete manner through manual inspection. Manual tracking is inevitable because (1) no indicators have so far been developed for evaluation by an automatic tracking system, (2) No automated system exists to evaluate the performance of GIS service provider and the client on the basis of performance indicators and (3) there is no centralized mechanism for penalizing negligence of either party. This paper proposes (1). a method to regulate the technical part of the GIS contract by suggesting a simple and wizard-based Graphical user interface. (2). Conversion of existing manually prepared contracts into electronic contracts through lexeme-based congregation which is done through labelled K Means Clustering. These converted clusters are then stored into centralized database. Back Propagation Neural Network (BPNN) is used to train the system on performance indicators defined for compliance by both contracting parties.

KEYWORDS: *Geographic information system, Back Propagation Neural Network, Clustering, Technical contracts, Contract, BPNN*

INTRODUCTION

Geographic Information System (GIS) is becoming popular in the industry which operates different networks like telecommunication, gas, and railways network. The networks of such industries are mapped on the base imagery by state of the art techniques of cartography and information system. For operations like capturing, storing, querying, analyzing and displaying the geographical data GIS are built. The design phase of GIS is distinct from conventional information system. It follows packaged approach, so squeezed that muffled licenses have evolved to complicated agreements (Kaminski & Perry, 2007) and contracts between client and service provider. GIS Contract has three main parts, first is service part in which the essential terms and conditions for delivering services to client and liabilities/obligations of both client and contractor are discussed. Technical part of the contract is its backbone and it covers the core scenarios involved in the design of GIS. Third is the resource part which is about the essential physical and human resources agreed between the parties in the light of technical and service conditions of the contract. This third part does not always become the part of GIS contract except in the case of where the requirements of GIS mapping center are to be sorted out (Longley *et al.,* 2011). GIS general contracts are in practice and are generally managed in both ways manually and electronically but mostly contracts are prepared and tracked manually in the progressing countries. Manual contracts are time consuming and less manageable. With the advent of technology, these manual contracts are now converted to electronic contracts by developing newer electronic Graphical User interfaces (GUIs) through which the data is entered to the database. A method to support direct conversion of these paper contracts to automated contracts without developing new GUIs is needed. A similar case in observed in a network-based company in Pakistan. An automated tracking module (ATM) can also become the part of this system which will track the progress of the contracting parties for which performance indicators are to be developed (LGO, 2013; Shaheen *et al.,* 2011b; Scottish Parliament report, 2018). In order to make new GIS contracts via wizard-based methods, there are two possibilities (1). To convert existing contracts to automated contracts (2). To develop new contracts. One of the techniques used for converting existing manually prepared contracts, includes usage of optical character recognition (Kotsiantis & Pinetelas, 2004; Coates *et al.,* 2011; Jain *et al.,* 1999). A new method based on pixel

Department of Software Engineering, Foundation University, Islamabad, Pakistan. E-mail: dr.shaheen@fui.edu.pk

statistics is proposed in which mean, standard deviation and entropy of pixel intensity is stored in a database. Labeled K Means clustering is applied on the dataset to classify the systematic diagrams of the contract in number of known clusters. Templates stored against each cluster are retrieved to assign values to the performance indicators of technical part of GIS contract. These performance indicators are also developed in this study. Backpropagation neural network (BPNN) is used to train the system on the performance indicators for generalized use in future (Shaheen *et al.,* 2011b; Jing *et al.,* 2012).

In extant literature, electronic contract management mainly discussed the transformation of manual contracts into electronic contracts. Various directions of electronic contract management are highlighted. For electronic contract preparation, a method was proposed by (Tan & Theon, 2000), contract binding and negotiation is proposed by (Perrin & Godart, 2004) and (Pong & Signgate, 2001) respectively. Moreover, the work on electronic contract management along with contract negotiation, execution and creation are discussed by (Pong & Signgate, 2001) and (Griffel *et al.,* 1998). An intelligent system for service part of contract is proposed by (Kaminski & Perry, 2007). In this paper, service level contracts and objectives are automated to track performance of contracting parties. Negotiations on service level Agreements (SLA) are managed through an intelligent agent. A lot of researchers believe that insertion of intelligent software agents, which are able to conciliate and commute in the situation, lead towards biased assessment of contracts (Dignum *et al.,* 2002; Pacheco & Carmo, 2003; Dellarocas, 2001). Message based collaboration is considered to be a better option to ensure a well managed contract. A frame work designed by (Abdel & Salle, 2002) is a source of message based collaboration between different stakeholders in an organization.

GIS contract includes three parts: (1) Service (2) Technical, and (3) Resource. Through layered and interactive architecture, a structure of GIS contract can be demonstrated. (Given in Fig. 1 and 2). All GIS three parts are shown in first layer of Fig. 1. The study is related to the technical part of GIS contract only. In GIS development technical steps which are undertaken are related to technical portion in a contract. The first and important step is to procure satellite images of area of interest (AOI) in order to build GIS. To hand

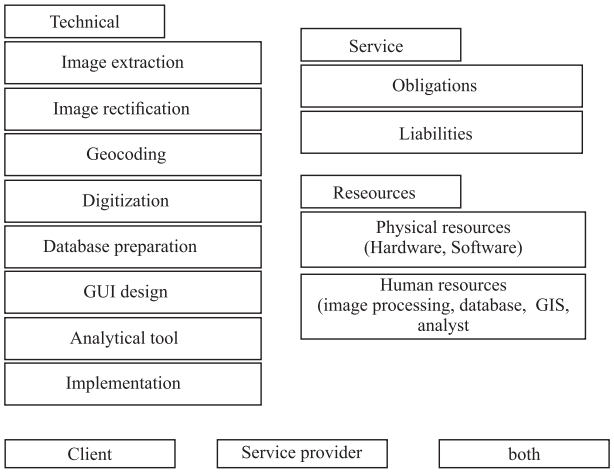

Fig. 1: Parts of a GIS Contract

it over to the service provider client is responsible to demarcate the area on map in this step. There can be different formats on which demarcation is done. For this purpose, some organizations use digital images whereas other may uses maps. The service provider rectifies the image, once an image is procured. The process in which image is converted on a real ground coordinated is called rectification, in the context of GIS (e.g. longitude and latitude (Longley *et al.,* 2011). Client remains a silent observer in this step. In some situations, client sometimes have to select projection system for rectification process. After that images which are rectified are then digitized. A process in which raster images are converted to vector form is known as Digitization (FangChih *et al.,* 2009). Network mapping in the form of engineering designs or diagrams are provided by client. Spatial and non-spatial data is generated by digitized ground-based map, which is in the form of satellite imagery (Shaheen *et al.,* 2013). Mostly Spatial data is handled and planed at service provider's end, while the client is supposed to propose attributes and database management system for non-spatial attribute data (Shaheen & Khan, 2015). After that, a database set is prepared successfully which is a new milestone for service provider and that's graphical user interface GUI.

This study proposes a method for automatic tracking of GIS contracts for which a method to convert existing signed contracts to automated modules by using artificial intelligence techniques is proposed and then method for tracking of the performance on the basis of performance indicators is given. Performance indicators for GIS contracts are proposed in section 2. Section 3 gives

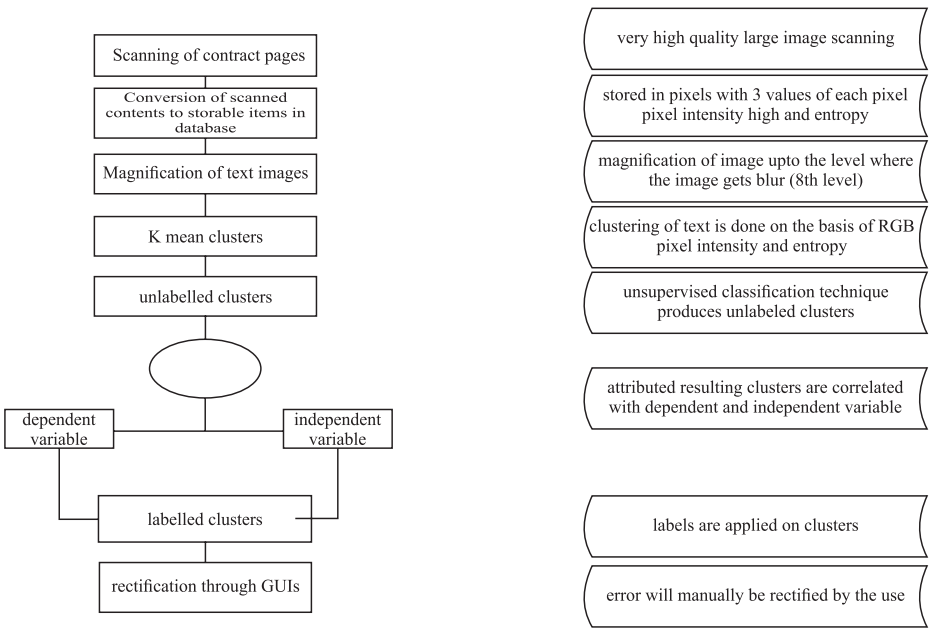

Fig. 2: Conversion of existing contracts

methodology for using Labeled K Means clustering and AI based conversion for tracking the contract. Section 4 discusses results/experiments of the study.

PERFORMANCE INDICATORS

In GIS contracts, Clients only determine technical logic and business logic of the required system. Conversion of pseudo code to a structured programmed module is a responsibility of service provider who serves as a consultant at this stage. An operational GIS system is developed, installed and implemented at client side through the above-mentioned steps. The implementation stage changes in different sets of situations.

Implementation of GIS can be done either by the distributed environment or as a desktop solution. The desktop solution is free of concurrency issues and also cost-effective. But the desktop solution doesn't usually satisfy technical requirements of the client. The distributed solution is either one browser-based or server based. Evaluation of these solutions is done on the basis of (1) Long-run benefits (2) Ease of management (3) Security and (4) Cost.

On the basis of these factors, performance indicators for implementation of GIS are given in Table 1. The performance indicator of resource term is overlapping with few of the performance measures in technical terms. In technical term contract, there are also some indicators that are also partially overlapping with each other. Performance measure criteria of the service provider and client can be qualitative and quantitative depending upon the type of GIS. For example, geocoding is measured by qualitative means whereas quantitative measurement can be defined for the evaluation of rectification step. (according to Table 1).

MATERIALS AND METHODS

Data mining is a domain that analyses historical data to ensure that the quality of knowledge derived from data is directly proportional to the amount of available data. A structured and modeled approach composed of fixed steps which organize the raw data into meaningful knowledge for example predictions are made and patterns are extracted from the data. Finding a likely grouping among different objects is said to be clustering. In data mining, methods of classification are categorized as: (1) Unsupervised Classification and (2) Supervised Classification. In supervised classification, the input data has the class labels, but in unsupervised classification the input data won't have the class labels (Shaheen *et al.,* 2010). A method of clustering these unsupervised datasets was proposed by (Shaheen *et al.,* 2011b) and is used in this study for conversion of existing signed

Table 1: Key Performance Indicators for GIS Contract Management

No	Name of indicator	Unit	C/S	No	Name of indicator	Unit	C/S
Procurement of Satellite Image							
1.	Demarcation of AOI	days	C	34.	Dependency of database	erd	S
2.	Procurement of satellite images	days	S	35.	Spatial database	days	S
3.	Resource person (AOI)	no.	C	36.	Attribute database	days	S
4.	Rate of image	/km2	S	37.	Penalty cost for delay	$/day	S/C
5.	3rd party?	boolean	S	38.	Penalty cost for error	$/day	S/C
6.	Reason for 3rd party	Text	S	Graphical User Interfaces			
7.	Resolution of image	no.	S	39.	No of fields	no.	S/C
8.	Penalty cost for delay	$/day	S/C	40.	No of images	no.	C
9.	Penalty cost for error	$/day	S/C	41.	No of prog modules	no.	S
Rectification							
10.	No of meetings with clients	no.	S	43.	GUI provision	days	S
11.	Projections	days	C	44.	GUI structure	days	C
12.	Base map	days	S	45.	GUI objects	days	C
13.	Extended documents	no.	C	46.	Integration of GUI	days	S
14.	Resource persons	no.	S/C	47.	No of GUIs	no.	S
15.	Penalty cost for delay	$/day	S/C	48.	Penalty cost for delay	$/day	S/C
16.	Penalty cost for error	$/day	S/C	49.	Penalty cost for error	$/day	S/C
				Analytical Modules			
Digitization				50.	No of analytical mod	no.	S
18.	No of digital maps	no.	C	51.	Lines of code	no.	S
19.	No of other docs	no.	C	52.	Method innovation	nominal	S
20.	Maps provision	days	C	53.	Code innovation	nominal	S
21.	Geocoding	days	S	54.	Involvement of 3rd party	boolean	S
22.	No of meetings	no.	S	55.	Resource person	no.	S
23.	Resource persons	no.	C	56.	Modules delivery	days	S
24.	Size of coded content	byte	S	57.	Integration of analytical modules	days	S
25.	Penalty cost for delay	$/day	S/C	58.	Penalty cost for delay	$/day	S/C
26.	Penalty cost for error	$/day	S/C	59.	Penalty cost for error	$/day	S/C
Database preparation				Implementation			
27.	No of database attributes	no.	S	60.	Type of implementation	nominal	S
28.	No of spatial attributes	no.	S	61.	No of systems	no.	S/C
29.	Time (spatial)	days	C	62.	No of privileges	no.	S
30.	Time (attribute data)	days	C	63.	Hardware cost	$	S
31.	Size of database	byte	S	64.	Software cost	$	S
32.	Relationship in database	no.	S	65.	Penalty cost for delay	$/day	S/C

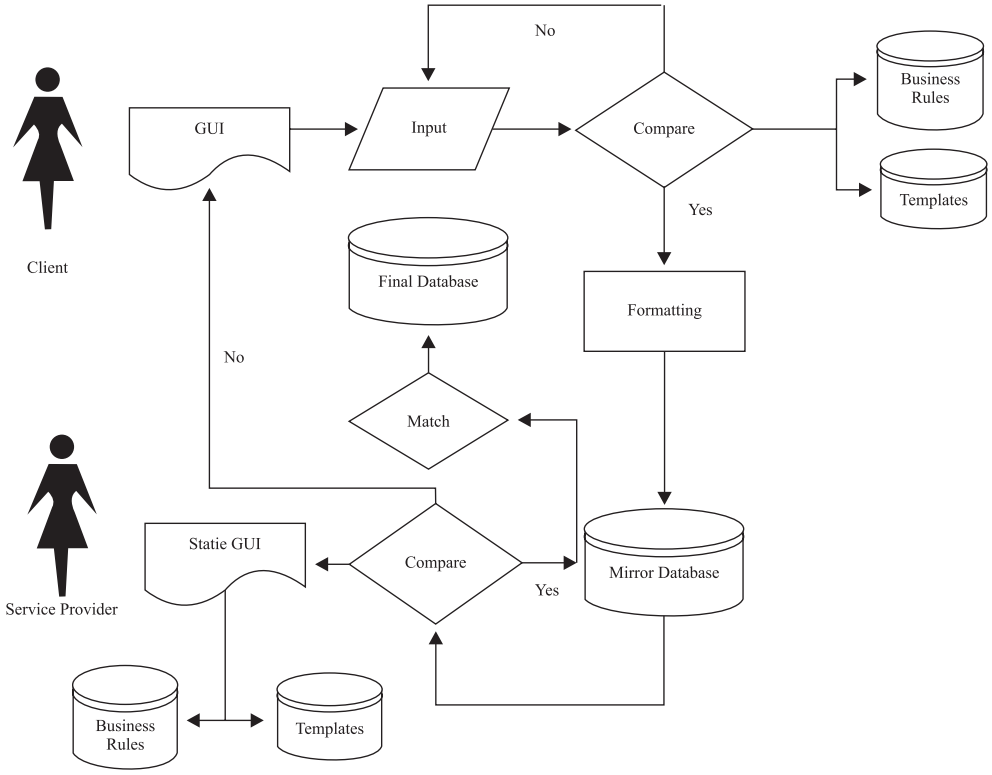

Fig. 3: Process flowchart

GIS contracts to automated modules.

Following are the steps for labeling clusters obtained by K Means clustering (Shaheen *et al.,* 2011b; Shaheen *et al.,* 2013b);

1. Given dataset is divided into a required number of clusters via K-Means Clustering technique.

2. From provided data set locate independent variables, so on the basis of these variables class values can be assigned.

3. Find association between independent variable and parameters of the dataset and assign same as weighted value.

4. Calculate the product of assigned value with actual dependent variable value.

5. Data points are assigned to the cluster with maximum association value based on frequent membership rule (Shaheen *et al.,* 2011b).

Conversion of Existing Signed Contracts

Optical character recognition (OCR) is applied to transform high-dpi scanned pages of existing signed contract to the text. The process is shown in Fig. 3. Some magnified images contain blur text but it does not really matter because of the post-processing applied on the images including magnification. Following values of magnified image are calculated for each scanned pixel (Shaheen *et al.,* 2013).

1. Pixels mean

2. Pixel standard deviation

3. Pixels entropy

These values are saved in database for applying K-Means clustering. Lexemes from the text are identified on the basis of clusters obtained. The unnamed cluster gives no information about the members of the cluster. Labeled clustering is used to name the lexemes identified by K Means method. At the end, prominent and important lexemes are obtained from the artifact

with the number of occurrences saved in the database. Stored datasets of lexemes are made available for users, service provider and client through GUI for manual verification. If there was some frequently used term in document different from existing keyword is classified by using correlation analysis between the term itself and the closest possible match. Equation 1 (correlation analysis) is used to calculate correlation between each lexeme and closest lexeme.

$$Corr\left(L_i, L_j\right) = \sum_{i=1}^{n} \frac{P\left(L_i, L_j\right)}{n} \quad (1)$$

Calculation of function P (Li, Nj) is described in Equation 2. In which Li defines occurrence of lexeme itself and Nj is occurrence of closest possible match.

$$P\left(L_i, L_j\right) = \frac{V\left(L_i, L_j\right) - V\left(L_i\right)V\left(L_j\right)}{V\left(L_i^{\,2}\right) - V^2\left(L_j\right)\sqrt{V\left(L_i^{\,2}\right) - V^2\left(L_j\right)}} \quad (2)$$

New Contract

The contract module is made up of validation control, template, GUI, and graphic panel. Strategies and business rules are provided by the clients and service providers as input through GUI which is then stored in a shared database. Mostly inputs have to be Boolean, numeric and in date/time format rather than strings or other formats. The business rules are checked against one another and if there exits some conflict, the notification is generated to the related party. Services are implemented in the system according to constraints and needs of the client. For example, if service provider tends to use some product which is not authorized by the client. Both the client and service provider were given the option to set standard values for performance indicators as given in Table 1, before signing electronic copy of the contract.

The system is trained on the terms and conditions set by performance indicators by using artificial neural network. Back propagation neural network (BPNN) is used in which major concern is to define transfer function and the principle of weight. BPNN is supervised learning algorithm which has constant activation function/network structure of the nodes rather it adjusts the weights in the system. Learning/ training of BPNN is done in two steps, i.e. forward pass and backward pass (Jing *et al.,* 2012).

In forward pass, it calculates total input of each node to give output by using activation function.

$$Out_k = f\left(\sum_{j=1}^{n} w_{ij}\, H_j\right) = f\left(\sum_{j=1}^{n} w_{ij}\, f\left(\sum_{j=1}^{n} w_{ij}\, X_j\right)\right) \quad (3)$$

$$(\mathrm{NI}) = \sum_{i,j-1}^{n} W_{ij} X_i \quad (4)$$

H is the output of the particular node and Wij is the adjusted weight.

Equation 3 and 4 are used to calculate output obtained at the output layer by giving the total input to the input layer which is then passed to the hidden layer whereas in second step backward pass, weights are adjusted according to a number of errors by sending back to the last hidden layer. Equation 5 adjusts the weight.

$$\Delta W_{jk} = -\xi \frac{\mathrm{dE}}{\mathrm{dW_{jk}}} \quad (5)$$

Where, ξ is the constant adjustment value.

This paper proposed a BPNN technique for understanding and learning performance indicator which is presented in table 1. It can keep checking the progress of both teams which should be according to the terms and conditions of the contract. BPNN had training sample which was presented to it. By evaluating the output value, the output of presented sample was calculated. Now calculated value was compared with the required value and after that the limit of error is calculated. Necessary alterations required to get output is done by measuring the scaling factor. Readjustment of a neuron to higher weight is done by sending back the local error in the last layer. This procedure will keep on repeating till the error becomes insignificantly small.

Experiments

Facts and data for testing the above procedure was collected from companies of Pakistan which hired the services of consulting companies for development of network GIS. Large size Cartographic maps were obtained from network or engineering drawings, i.e. 580 engineering design sheets were presented to show distribution of a network of 1 city of a company. Lexemes were made from contract document which was scanned and then parsed. Out of 82 tokens only 62 were converted into understandable lexemes whereas other tokens were not understandable. Service provider and client used shared database which is also known as the spatial database

because it was previously hosting GIS applications. Table 2 presents determined value of some correlated matrices of lexemes. Table 3 and Fig. 5 give the results of clustering of lexemes from documents which were scanned through WEKA.

Lexemes were attached with GUI after extraction from documents. Lexemes were verified and information was extracted (as it was not derivable from lexemes) by the user with the help of provided GUI. There was Graphical panels and the Embedded GUIs systems associated with digitization performance indicators i.e. GUI formats and fields, GPS data collection, digitization rules and codes, database resources, projections and rectification scale, database attributes/relationships, implementation and satellite image procurement. Table 4 presents results of proceed lexemes which were manually changed by user.

REFERENCES

1. *Kaminski, H., & Perry M., (2007), "Employing intelligent agents to automate SLA creation", CesarePautasso and ChristophBussler, pp. 33-46.*

2. *Longley, p., Michael, F., Goodchild, D., & Maguire, D., & Rhind, D, (2011), Geographic Information System and Science, 3rd Ed, New York, USA.*

3. *Local Government Association, (2013), "Making Saving from Contract Management", A report by Local Government Association London.*

4. *Shaheen, M., Aslam, M., Shahbaz, M., Khan, J., & Shaheen, N., (2011), "An Intelligent Mechanism for GIS contracts automation", in World Congress on Engineering and Computer Science, International Conference on Data Mining and Knowledge Engineering, USA.*

5. *Contractor performance management strategy. Website of The Scottish Parliament," 2018. [Online]. Website: http://www.parliament.scot/abouttheparliament/65849.aspx. Accessed on: Oct 18, 2018.*

6. *Kotsiantis, S., & Pintelas, P., (2004), "Recent advances in clustering: A brief survey", WSEAS Trans. Inform. ScAppl, pp. 73-81.*

7. *Coates, A., Carpenter, B., Case, C., Satheesh, S., Suresh, B., Wang T., J. Wu, D., & Y. Ng, A., (2011), "Text detection and character recognition in scene images with unsupervised feature learning", in International Conference on Document Analysis and Recognition.*

8. *Jain, A., Murty, M., & Flynn P., (1999), "Data Clustering: A Review", ACM ComputSurv: Vol 31, pp. 264-323.*

9. *Jing, L., Ji-hang, C., & Jing-yuhan, S., & Fei, H., (2012), "Brief introduction of backpropagation neural network algorithm and its improvements," AdvIntell Soft Comput; Vol 69, pp. 553-558.*

10. *Tan, Y., & Theon, W., (2000), "DocLog: An electronic contract representation language", in 11th International Workshop on Database and Expert System Applications; London, UK.*

11. *Perrin, O., & Godart, C., (2004), "An approach to implement contracts as trusted intermediaries", in First IEEE International Workshop on Electronic Contracting.*

12. *Pong, M., & SignGate, E., (2001), "Electronic contract signing gateway", in 25th Annual International Computer Software and Applications Conference.*

13. *Griffel, F., Boger, M., Weinreich, H., Lamersdorf, W., & Merz, M., (1998), "Electronic contracting with COSMOS – how to establish, negotiate and execute electronic contracts on the Internet", in International Enterprise Distributed Object Computing Workshop.*

14. *Kwok, T., & Nguyen, T., (2006), "An enterprise electronic contract management system using dual XML and secure PDF documents", in 10th IEEE Intl Enterprise Distributed Object Computing Conference Workshops.*

15. *Dignum, V., Meyer, J., & Weigand, H., (2002), "Towards an organizational-oriented model for agent societies using contracts", in First International Joint Conference on Autonomous Agents and Multiagent Systems, ACM.*

16. *Pacheco, O., Carmo, J., (2003) "A role based model of normative specification of organized collective agency and agents interaction", Auton Agents &MultiagentSyst, p. 145–184.*

17. *C. Dellarocas, (2001), "Negotiated shared context and social control in open multi-agent systems", R. Conte and C. Dellarocas, ed, Social Order in MAS. Kluwer.*

18. *Abdel, B., & Salle, M., (2002), "Integrated Contract Management, 9th workshop of HP Openview University Association", pp. 4-13.*

19. *FangChih, L., Tzu-Ying, L., & Yeng-Hun, L., (2009), "Maps and GIS Digitization Procedure Guides", International Collaboration and Promotion of Taiwan E-Learning and Digital Archives Program, pp. 18-42.*

20. *Shaheen, M., Shahbaz, M., & Guergachi, A., (2013), "Context based positive and negative spatio temporal association rule mining", Knowl-Based Syst, pp. 261-273.*

21. *Shaheen, M., & Khan, Z., (2015), "A method of data Mining for selection of site for wind turbines", Renew SustEnerg Rev 2015.*

22. *Shaheen, M., Shahbaz, M., Guergachi, A., & Khan, Z., (2010), "Data mining applications in hydrocarbon exploration", ArtifIntell Rev, pp. 1-18.*

23. *Shaheen, M., Shahbaz, M., Guergachi, A., Khan, Z., (2011) b "Mining sustainability indicators to classify hydrocarbon development", Knowl-Based Syst, p. 1159–1168.*

24. *Shaheen, M., Iqbal, S., & Basit, F., (2013) "Labeled Clustering: A unique method to label unsupervised classes", 8th International Conference on Internet and Secured Transaction, pp. 210-214.*