Analytical Study and Performance Comparison of Various Machine Learning Tools

Sonu Bala Garg[1], Jatinder Garg[2*]

[1]IK Gujral Punjab Technical University, Jalandhar (Punjab)

[2]Baba Hira Singh Bhattal Institute of Engineering and Technology, Lehragaga (Punjab)

**Corresponding Author: jatindergarg@yahoo.com*

ABSTRACT

In today's modern world, data has been increasing enormously in volume, velocity and variety. It is difficult to process enormous amount of data using traditional data processing techniques. This had lead to the emergence of a new technology named data mining. Data mining is the process of extracting hidden data, patterns and trends from large amount of data stored at data warehouse and various databases. It has various mining techniques like clustering, prediction, classification, association and regression to extract out important information from the given data. The objective of this paper is to evaluate some of the available data mining tools based upon their features, aim, requirements, algorithm supported and platform supported. It will help the users to select the best tool for their application. The analysis has shown that the Rapid miner and Mallet tool support classification, KNIME and Orange tools supports clustering and SSDT, WEKA, Apache Mahout and Oracle tools support both clustering and classification.

Keywords: Data mining; Knowledge discovery in databases (KDD), data mining tools

1. INTRODUCTION

In the modern world, most of the things are being processed by computers; as a result enormous amount of data are generated daily. The data is rapidly growing day by day, these data are not useful as long as they are not analysed and converted to a human understandable form [1]. Data can be static or dynamic. Static data is easy to handle and processed therefore, dynamic data is little bit difficult because it refers to high voluminous and continuously changing. Dynamic data changes with time that is why it is difficult to manage. Data can be in any form like sequential, audio signal, video signal, spatial temporal, temporal, time series etc [2]. Data mining is the process of extracting hidden values, patterns and information from existing data which are stored in data warehouse or databases. In data mining, firs of all the data is extract, then cleaned to make it usable by removing unnecessary noisy and missing values. It is then processed by using various tools and techniques.

Data mining has been used in various fields such as manufacturing, marketing, agriculture, chemical, aerospace, stock market etc. to enhance the business. In almost all such fields it has given out magnificent results, thereby making it a widely used method

of data analysis throughout the world. A wide variety of data mining techniques are available to analyse the data of interest. These include techniques like clustering, classification and association rule mining and others. Some of the commonly used data mining techniques have been shown in Fig. 1.

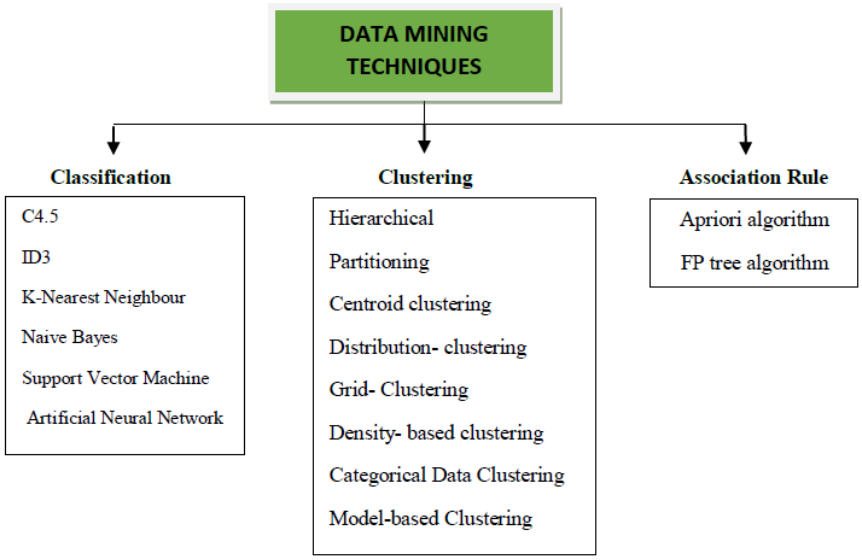

Figure 1: Commonly used data mining techniques

A. KDD (Knowledge Discovery in Databases)

The core part of the data mining is known as KDD (Knowledge Discovery in Databases). It is an iterative process in which interpretation measures can be improved, new data integrated and transformed to produce more accurate results. It involves using some essential steps such as data selection, data cleaning, data transformation, pattern recognition, Data presentation & Interpretation and data evaluation [3].

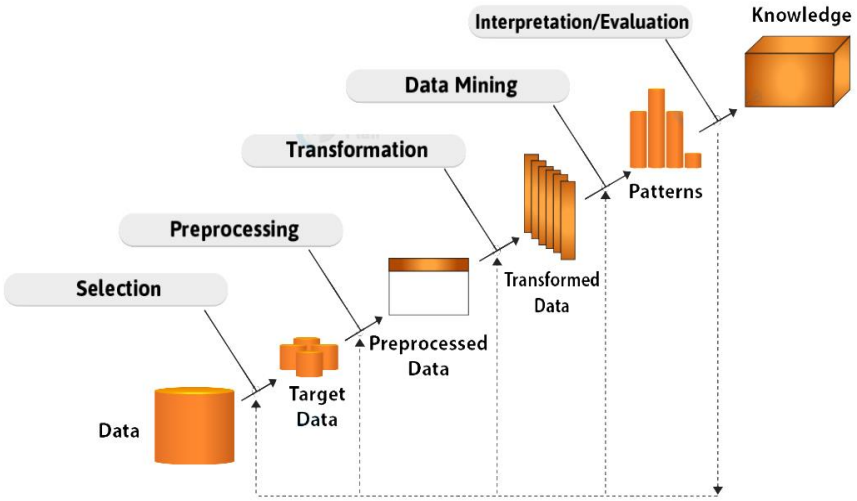

Figure 2: KDD (Knowledge discovery in databases process)

Various steps of the KDD process and their function are described below:

1. Data Cleaning: It is used to remove noisy and irrelevant data from the dataset.
2. Data Integration: Different types of data combined together in data warehouse from multiple sources.
3. Data Selection: It is a collection of relevant data for analysis and is called data selection.
4. Data Transformation: It is used to transform the data into appropriate form.
5. Data Mining: It is a process of extracting useful patterns and information by using relevant techniques.
6. Pattern Evaluation: It is used for summarization and visualization to make the data user understandable.
7. Knowledge Representation: Results are visualized in the form of reports, tables, charts and pattern. This step is called knowledge representation.

B. KDD (Knowledge Discovery in Databases)

Data mining provide us a variety of techniques for data analysis and pattern analysis, such as clustering, classification, and association rule for data manipulation.

1) Classification: Classification is one of the data mining technique that is used for processing of structured or unstructured data. The main goal of classification is to distinguish between data into a number of classes that fall under the same category. There are various types of classification algorithms, which are used for the classification of data mining [4], as given below:

a) C4.5

b) ID3

c) K-Nearest Neighbour classifier

d) Naive Bayes

e) Support Vector Machine

f) Artificial Neural Network election

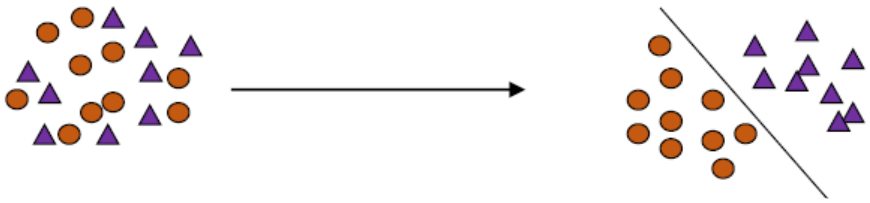

Figure 3: Representation of classification technique

2) Clustering: Clustering is one of the most important techniques of data mining which is widely used in various fields of research study like data mining, statistical data analysis, Machine learning, Biology, Data pattern recognition, Image analysis, and Information retrieval. Clustering is the process of grouping of similar data items into a set of clusters.

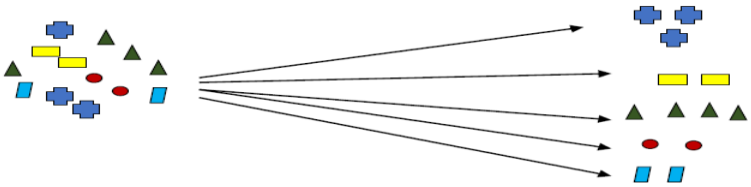

Figure 4: Representation of clustering technique

There are many clustering techniques as follows *[5]:*

a) Hierarchical clustering

b) Partitioning clustering

c) Centroid based clustering

d) Distribution-based clustering

e) Machine Grid-Based Clustering

f) Density-based clustering

g) Categorical Data Clustering

h) Model-Based Clustering C4.5

3) Association Rule: Association rule mining is a data mining technique broadly used knowledge discovery technique basically used for finding the frequent patterns, correlations between data, associations or structures among the sets of items or objects in database. Association rule technique having two main important properties namely Support and Confidence as shown below:

Support (AB) = P (AUB)
Confidence (AB) = P (B|A)

There are many Association rules mining algorithms [6] such as:

a) Apriori algorithm

b) FP-tree algorithm ID3

2. DATA MINING TOOLS

A brief description of various commonly used data mining tools is given under:-

A. Rapid miner

It is a very robust and very powerful Java based data mining tool developed by Rapid miner company and it is freely available for use. It provides user friendly integrated environment and is used for various applications like education, business, research, training, and various machine learning tasks like pre-processing, predictive analytics, modeling, visualization etc. It is used for extraction, transformation and loading of data [7].

B. ORANGE

Orange is an open source component-based tool which is used for data analysis, classification and visualization purpose and is developed at the Bioinformatics laboratory by the faculty of Computer & Information Science at the University of Ljubljana, Slovenia. It is used in mining of data through python scripting or visualization of programs. It is a Python library used for data manipulation and widget alteration. The main features of orange are data visualization such as bar graphs, trees, scatter plots, dendrograms, heat maps etc. Regression method is also being used in Orange where ensembles are basically wrappers around learners [8].

C. WEKA

WEKA (Waikato Environment for Knowledge Analysis) is open source JAVA based software tool, developed by the University of Waikato in New Zealand, issued under GNU General Public License, is freely available for use. It can be used to solve the real life problems. It is a collection of visualisation tools and various data mining tasks like for classification, clustering, and association rule extraction etc. WEKA is used for several different tasks as given below:

Explorer: It gives an environment for explore specific information from a given dataset.

Experimenter: Which provides users / researchers with the possibility of performs experiments and statistical tests between learning schemes.

Knowledge flow: It facilitates users with the same services as an explorer but with a drag-and-drop interface. This too supports incremental learning.

Simple CLI: It provides users with a simple command line interface and selecting this option from WEKA allows you to run WEKA commands directly [9].

D. KNIME

KNIME (Konstanz Information Miner) is a open source JAVA based language developed by a team of software engineers at University of Konstanz and more efficient data mining tool used for data analysis (Extraction, transformation and loading of data). It is used to perform various data mining tasks like data pre-processing, classification, clustering. and is also used in Business intelligence, customer relation management data analysis, financial data analysis, pharmaceutical research etc. [10].

E. Tanagra

Tanagra is a more powerful free machine learning software developed by Ricco Rakotomalala at the Lumière University. It is written in C, C++ and Java. It is mainly used for academic and research purposes. It is widely used for various data mining tasks such as predictive analysis, classification and regression, statistical learning, machine learning, association rule learning, machine learning etc. It follows the following steps: Data extraction, Data transformation, Data analysis, Data visualisation, Data conversion, Data cleaning. In comparison to Weka, Tanagra has an easier to use Interface [11].

F. SQL server data tools (SSDT)

It stands for SQL server data tools. It is a licensed product developed by SAS Institute. It is highly scalable, provides graphical user interface and is written in C, C++. It is widely used to build, maintain, debug and refactor databases. SAS firstly mine the data and then manage the data from different sources. It gives the opportunity to user for analyse big data.

G. Apache Mahout

Apache Mahout is open source software, written in java and Scala, developed by Apache Software Foundation. It is mainly used when dataset is very large, too large to process on single machine. The main aim of Apache Mahout is creating extensible machine learning algorithms. It performs popular machine learning algorithms such as clustering, classification, finding similarities from large datasets. It works in distributed environment [12].

H. Oracle Data Mining (ODM)

Oracle data mining is Proprietary Licensed software developed by Oracle Corporation that provides powerful data mining algorithms to discover new patterns from hidden data. It is written in assembly language, C, C++. It runs on Windows & Linux operating system. It is more reliable and more powerful application. It is a part of oracle relational database ODM, has several data mining & analysis algorithms i.e. prediction, regression, classification, association, anomaly detection, feature selection. In ODM, models are stored in the database as database objects and implemented in oracle database kernel [13].

I. Rattle

Rattle means “R Analytical Tool to Learn Easily”, freely available and open source (GNU general public license) software use R language developed by Graham Williams. It is popular and provides graphical user interface for data mining. It presents the data in the form of statistical and visual so that it can be easily understandable. Rattle contains multitude of R packages that are necessary for the data mining [14].

J. DataMelt

DataMelt is also known as DMelt, is an open source developed by DataMelt community Led by S.Chekanov. It written in JAVA language and multiplatform utility, the program can run on Windows, Linux, Mac operating systems. It has distributed memory processing and highly scalable. It provides computational and visualization environment, and mainly used for big data analysis. It creates high quality of graphics and images. It draws 2D, 3D plots [15].

K. SAS Data Mining

SAS Data Mining is a Proprietary Licensed developed by SAS Institute at North

Carolina State University for statistical analysis. SAS is software that can extract, modify, manage and extract data from various sources and perform statistical analysis. The predictive analysis allows the users to learn future from the past [15].

L. R

R is a free programming language tool written in C++ developed by R Core Team. It is primarily used for machine learning, data mining, statistical computing and graphics. It can be used for various statistical tests modeling, data analysis, and various data mining tasks like classification, clustering etc. The main feature of R is its ease of use. Further, it is designed for graphics, formulae and mathematical symbols. It handles data very effectively and provides better storage facility [15].

3. COMPARISON OF VARIOUS DATA MINING TOOLS

The comparison of various data mining tools on the basis of different parameters has been performed [16-20] and has been presented in tabular form below:-

Table 1. Comparison of data mining tools

Tools	Website	Developer	License Type
Rapid Miner	https://rapidminer.com/	Rapid Miner	Open source
Orange	http://orange.biolab.si/	University of Ljubljana, Slovenia	Open source
WEKA	www.cs.waikato.ac.nz/ml/weka	University of Waikato New Zealand	Free software
KNIME	http://www.knime.org/	Team of software engineers at University of Konstanz as a proprietary product	Open source
Tanagra	http://eric.univlyon2.fr/~ricco/tanagra/en/tanagra.html	Ricco Rakotomalala at the Lumière University	Free Open source
SSDT	https://docs.microsoft.com/en-us/sql/ssdt/download-sql-server-data-tools-ssdt?view=sql-server-	By Microsoft	Licensed
Apache Mahout	https://mahout.apache.org/	Apache Software Foundation	Open source
Oracle Data Mining	https://www.oracle.com/	Oracle Corporation	Proprietary License
Rattle	https://rattle.togaware.com/	Graham Williams	Open source

DataMelt	https://jwork.org/dmelt/	DataMelt community Led by S.Chekanov	Open source
SAS Data Mining	https://www.sas.com/	SAS Institute	Proprietary Licensed
R	http://cran.r-project.org	Ross Ihaka and Robert Gentleman	Free Software

Table 2. Comparisons of tools on the basis of language and file type supported

Tools	Language supported	File type Supported
Rapid Miner	JAVA Language	accdb, Arff, csv , dbf , dta, hyper, mdb, qvx, sas, sav , tde, xls/xlsx, xml, xrff
Orange	Python computing language	xls/xlsx, csv, txt
WEKA	JAVA Programming Language	Arff, arff.gz, bsi, csv, dat, data ,json ,json.gz ,libsvm, m,names, xrff ,xrff.gz
KNIME	JAVA Programming Language	pdf, Docx, Doc, PubMed, dml, xls/xlsx, csv
Tanagra	C, C++, Java	xls/xlsx
SSDT (SQL Server Data Tools)	C, C++	NTFS or ReFS file formats
Apache Mahout	Java, Scala	Apache Spark, H2O, and Apache Flink
Oracle Data Mining	Assembly language, C, C++	xls/xlsx
Rattle	R	csv, txt, excel, Arff, odbc, R Dataset, RData File, Library Packages Datasets, Corpus, and Scripts
DataMelt	Java, Jython	scripting languages such as Jython (Python), Groovy, JRuby, BeanShell.
SAS Data Mining	C programming language	such as gif, jpg, pdf, power point, and word, csv, xml, url , html files
R	C, C++ and Fortran	txt, csv or excel file

Table 3. Comparison of various tools on the basis of their aim and platform supported

Tools	Aim	Platform supported
Rapid Miner	Deep learning	Windows, Linux, Mac OS
Orange	Machine learning & Data mining	Windows, Linux, Mac OS
WEKA	General ML package	Windows, Linux, Mac OS
KNIME	Data Pre-processing	Windows, Linux, Mac OS
Tanagra	for Academic & Research purpose	Windows
SSDT (SQL Server Data Tools)	to build, maintain, debug and refactor databases	Windows
Apache Mahout	creating machine learning algorithms	Linux, Apple OS
Oracle Data Mining	provides excellent data mining algorithms	Windows, Linux
Rattle	Exposes the statistical power	GNU/Linux, Macintosh OS/X, and MS/Windows
DataMelt	data analysis and visualization	Windows, Linux, Mac OS and Android operating system
SAS Data Mining	for statistical analysis & data management	Windows, Linux, Unix, CentOS, Apple OS
R	Statistical computing & graphics	Windows, Mac OS

Table 4. Comparison of various tools on the basis of features and algorithm supported

Tools	Features	Algorithm Supported
Rapid Miner	Graphical user interface Analysis processes design Multiple data management methods Data from file, database, web, and cloud services In-memory, in-database and in-Hadoop analytics Application templates, D graphs, scatter matrices, self-organizing map	Logistic Regression, Rigde Regression, LARS, Decision.

	GUI or batch processing Integrates with in-house databases Interactive, sharable dashboards	
Orange	Graphical user interface, interactive data visualization, improved data pre-processing	Linkage-based, k-means, ANN based Self organizing map, Partition around Medoids, fuzzy c-means clustering
WEKA	GUI, Data Pre-processing, classification, Regression, Clustering, Association rule and data Visualization	Linkage-based, k-means, X-means, EM, DBSCAN, OPTICS
KNIME	Big Data Extensions, Data Blending, Tool blending, Metanode linking, Local automation, Workflow difference, powerful analytics	Linkage-based, k-means, fuzzy c-means, X-means(E), EM(E), DBSCAN(E), ANN based(E), OPTICS(E)
Tanagra	Easy to use data mining software, Interactive utilization, A wide set of data sources, Data cleansing, Direct access to data warehouses and databases.	Supervised learning algorithms, own algorithms
SSDT (SQL Server Data Tools)	Declarative schema-based design, Data comparison features, T-SQL code editing and debugging, Support for MS-SQL Server 2005 and later	Decision Tree algorithm, Decision Trees, Naïve Bayes, Neural network, Linear regression
Apache Mahout	Collaborative filtering, Clustering, Classification, Frequent item set timing, Distributed Algebraic optimizer, Linear algebra operations	Naive Bayes Implementations, Random Forest, Canopy Clustering, k-Means Clustering, Fuzzy k-Means, Streaming k-Means, Spectral Clustering
Oracle Data Mining	Data transformation and model analysis, Anomaly detection, Classification, Regression,	Decision Tree (DT), Generalized Linear

	Feature selection, Clustering, Feature extraction, Text and spatial mining	Models (GLM), Minimum Description Length (MDL), Naive Bayes (NB), Support Vector Machine (SVM)
Rattle	Features statistical tests, Clustering, Modeling, Evaluation, Transformations, Visualization	Neural network, Regression, Decision Tree, Random Forest.
DataMelt	Access to java API, Access to image gallery with code examples, Web access, community forum and bug tracker, Online manual	Cluster analysis (K-means clustering analysis (single and multi-pass), Fuzzy (C-means) algorithm, agglomerative hierarchical clustering)
SAS Data Mining	Easy to use GUI and batch processing, Advanced predictive and descriptive modeling, high performance capabilities, scalable processing, open source integration with R.	Naïve Bayes Classifier Algorithm, K Means Clustering Algorithm, Support Vector Machine Algorithm, Linear Regression, Artificial Neural Networks, Nearest Neighbours, Decision Tree
R	Graphics Visualization, Spatial data analysis, Clustering, Text mining, Statistics, Graphics, Data manipulation.	Linkage-based, k-means, AGNES(E), BIRCH(E), X-means(E), EM(E), DBSCAN(E), ANN based(E), Fuzzy neural networks

4. CONCLUSION

In this paper, a brief description of various commonly used data mining tools has been presented. Each tool has its own features, advantages and limitations. Rapid Miner, WEKA, KNIME, Mallet, DataMelt supports JAVA Language; while Tanagra, Orange, SAS and R data mining supports C, C++; similarly Orange and DataMelt supports python language. All the tools do not support all the data mining operations. In the last section, detailed comparison of these data mining tools has been done and the results have been presented in tabular form for the readers. This comparison will help the researchers focus

on the different issues of data mining and will help them to choose the right tool for the given task. There are several other data mining tools which are available and can also be use to perform different data mining operations, but these have not been covered in this study due to time and resources limitations. The future scope of this paper is to study those data mining tools with the aim of presenting a more comprehensive comparison.

REFERENCES

[1] M. Bharati and M. Ramageri, "Data mining techniques and applications," *Indian Journal of Computer Science and Engineering*, vol. 1, 2010.

[2] M. Phridvi Raj and C. Guru Rao, "Data mining–past, present and future–a typical survey on data streams," *Procedia Technology*, vol. 12, pp. 255-263, 2014.

[3] S. B. Garg, A. K. Mahajan, and T. Kamal, "An Approach for Diabetes Detection using Data Mining Classification Techniques," *Research Cell: An International Journal of Engineering Sciences*, vol. 26, pp. 202-218, 2017.

[4] S. K. Sarangi, V. Jaglan, and Y. Dash, "A Review of Clustering and Classification Techniques in Data Mining," *In ternational Journal of Engineering, Business and Enterprise Applications*, vol. 13, pp. 140-145, 2013.

[5] D. Patel, R. Modi, and K. Sarvakar, "A Comparative Study of Clustering Data Mining: Techniques and Research Challenges," *International Journal of Latest Technology in Engineering, Management & Applied Science*, vol. III, pp. 67-70, 2014.

[6] A. R. Kulkarni and S. D. Mundhe, "Data Mining Technique: An Implementation of Association Rule Mining in Healthcare," *International Advanced Research Journal in Science, Engineering and Technology*, vol. 4, pp. 76-85, 2017.

[7] Galvanize. (2018, 15 January 2018), *4 Data Mining Techniques for Businesses (That Everyone Should Know)*. Available: https://blog.galvanize.com/four-data-mining-techniques-for-businesses-that-everyone-should-know/

[8] M. Kukasvadiya and N. Divecha, "Analysis of data using data mining tool orange," *Int. J. Eng. Develop. Res*, vol. 5, pp. 1836-1840, 2017.

[9] M. Gera and S. Goel, "Data mining-techniques, methods and algorithms: A review on tools and their validity," *International Journal of Computer Applications*, vol. 113, pp. 22-29, 2015.

[10] (2018, 15 January 2018). *KNIME*. Available: https://en.wikipedia.org/wiki/KNIME

[11] Wikipedia. (2018, 15 January 2018). *Tanagra (machine Learning)*. Available: https://en.wikipedia.org/wiki/Tanagra_(machine_learning)

[12] S. Schelter and S. Owen, "Collaborative filtering with apache mahout," presented at the ACM RecSys Challenge, Dublin, Ireland., 2012.

[13] (2018, 15 January 2018). *Introduction to Oracle Data Mining*. Available: https://web.stanford.edu/dept/itss/docs/oracle/10gR2/datamine.102/b14339/1intro.htm

[14] G. J. Williams, "Rattle: a data mining GUI for R," *The R Journal*, vol. 1, pp. 45-55, 2009.

[15] D. Shah. (2017, 15 January 2018). *Data Mining Tools*. Available: https://towardsdatascience.com/data-mining-tools-f701645e0f4c

[16] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, *et al.*, "KNIME-the Konstanz information miner: version 2.0 and beyond," *AcM SIGKDD explorations Newsletter*, vol. 11, pp. 26-31, 2009.

[17] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2016.

[18] Q. A. Al-Radaideh, E. M. Al-Shawakfa, and M. I. Al-Najjar, "Mining student data using decision trees," in *International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan*, 2006.

[19] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping multidimensional data*, ed: Springer, 2006, pp. 25-71.

[20] C. C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," in *Mining text data*, ed: Springer, 2012, pp. 77-128.