International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-6, Issue-8)

Research Article | **August 2017**

Development of BPMN Architecture using Integration of Data Mining

Ritu Saluja
Research Scholar, Computer Science Engg., Galgotias University, Noida, Uttar Pradesh, India

Dr. S. K. Singh
Professor, Computer Science Engg. Galgotias University, Noida, Uttar Pradesh, India

Dr. A. K. Chaturvedi
Professor, Computer Science Engg. JIMS, Noida, Uttar Pradesh, India

Abstract-

Data mining is an important factor of success for modern business processes. Modern Integrating data mining with business solutions will improve the business process model (BPM) and enhance the overall business process management framework. The proposed architecture promises to provide great support for flexible design, deployment and managementof business processes. Incorporating services of data mining in order to choose the appropriate business model, determining missing standards and deploying machine learning techniques in an applicable manner is a challenging task discussed in the paper. Theproposed work describes the overall contribution of data mining services and validates the novelty in architecture by defining user roles for business, decision mining and IT standards. Supervised and unsupervised learning technique is used for determining similarity between the proposed model and the previously stored models.

Keywords- Data Mining, Business Processes, Integration, Business Process Model Notations (BPMN),Machine Learning, supervised Learning, Clustering.

I. INTRODUCTION

More and more people, both in industry and academia, consider Business process model or process mining, as one of the most innovations in the field of business processes and machine learning on the other. Business Process Modeling Notation (BPMN) is a flow chart method that models the steps of a planned business process from end to end. A key to Business Process Management, it visually depicts a detailed sequence of business activities and information flows needed to complete a process.the automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules**.**Due to the rapid growth in the field of data mining and advancement in the Information Technology, process mining can be achieved by integrating business process models with data mining techniques. With the increase in the competition in the market, business process modeling is adapter in which a new process model is created or an existing process model is adapted in redesigning phase. Our objective is to use some goal is to integrate a data mining process into this business process in order to improve the business. This has been justified by Dennis Wegener [1] and A. Rozinat [2] who surveyed on the importance of the business process intelligence by incorporating techniques of data mining into BPMN. Therefore tools have to be developed for predictive modeling and development of various BPMN models [3].Current research progress in business process management (BPM) is based on defining and managing business processes in BPEL- and BPMN-based [4,5]service-oriented environments. Such modern BPM environments provide flexibledesign, management and deployment of business processes. Given that thesepowerful BPM environments and the CRISP model exist, one could assume that it is very straightforward to efficiently integrate data mining in business processes.But is this really the case? In this paper we make the point that inpractice still many redundancies and inefficiencies exist. We believe that the integration of data mining into business processes in such environments can be simplified by developing an integrated approach. We aim at a concept for the integration that facilitates the modelling of the data mining process within thebusiness process as well as the technical deployment into the business IT environment.This includes an integrated view on the interfaces between data miningand the business, as well as additional specifications, standards and definitions.

The goals of the paper are as follows:

- To define and explain the related terms in business process modeling.
- To explicate the applications of data mining techniques.
- To propose the architecture by incorporating techniques of data mining.
- To evaluate the techniques used in integration.

The paper has been organized as follows. Section 1 of the paper introduces the concept of BPMN architecture. Section 2 discusses the background. Section 3 discusses the data mining techniques. Section 4 explains the proposed architecture, Section 5 shows the analyses and evaluatesthe results and the last section concludes.

II. BACKGROUND

In this secondary study carried out by Dennis Wegener, it was pointed that a concept for theintegration that facilitates the modelling of the data mining process within the business process as well as the technical deployment into the business is a crucial factor of success for business today [1].Business process management (BPM) is a discipline combiningon Integrating Data Mining into Business Processes and software capabilities and business expertise to accelerate business process improvementand to facilitate business innovation [6]. Integrating data mining in business processes involves different groups of users with different responsibilities,knowledge and background. Naively speaking, this will include the roles ofbusiness experts, who are in charge of the business but do not have knowledgein the implementation and in data mining, data miners, who are experts in datamining and who are in charge of designing and implementing the data miningprocess, and the IT experts, who manage the hardware and software resources ofthe business and who are in charge of implementing and integrating the businessprocess in the IT environment.

A. Business Process Model

Business process models can also be represented in BPMN[7] which also contains guidelines on their transformationinto BPEL. The intent of the BPMN for business process modeling is very similar to the intent of the Unified Modeling Language for object-oriented design and analysis. To identify the best practices of existing approaches and to combine them into a new, widely accepted language. The set of ancestors of BPMN includes graph-based and Petri-net-based process modeling languages, such as UML activity diagrams and event-driven process chains [8]. The BPMN standard defines a notation and a metamodel that organizes the concepts used in the notation. The graphical notation of a business process is complemented with a set of attributes. These attributes can be associated with the complete process diagram and with particular elements. Some attribute values have implications on the visual appearance of the symbols used in process diagrams. For instance, whenever a gateway activates a single outgoing edge from a set of outgoing edges, the gateway is marked with the X symbol to indicate its exclusive or split semantics. The business model can be collaborated with a layer located between the application layer and the middleware layer so it can helps the software developers to understand the process flow and also help to find a point of understanding between programmers and business analyst's point [9]. There are primarily three elements in a CMfg system which are:

- The providers which own and provide the manufacturing resources and provide cloud manufacturing service.
- The operators which operate the CMfg platform to deliver services and functions to providers, consumers, and thirdparties.
- The consumers which use the manufacturing cloud services available in a CMfg service platform, consumers are usuallysmall and medium enterprises [6].

1) Life Cycle

The business process lifecycle includes the following phases explained in Fig 1.

- Design phase: the process is designed
- Configuration phase: model coded into conventional software.
- Enactment (execution) / monitoring phase: process running and monitored by management, to see if changes are needed.
- Adjustment phase: changes made according to the previous phase.
- Diagnosis/requirements phase: evaluates the process and monitors new requirements (new policies, laws, etc.)

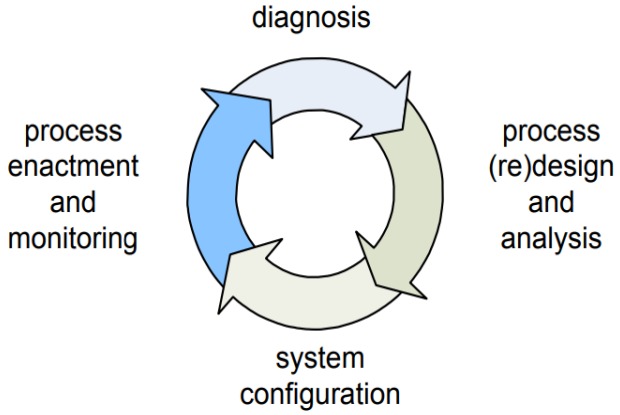

Fig 1: Lifecycle of Business Process

B. Data mining

Data mining means extracting (unknown) patterns from data. In general, a data mining processincludes several iterations of single data mining steps (algorithm executions).The goals of the data mining process are defined by the intended use of thesystem from the user perspective and can be classified into two types: verification,where the system is limited to verifying the user's hypothesis, and discovery, where the system autonomously finds new patterns. The discovery goal can befurther subdivided into prediction, where the system finds patterns for predicting the future behaviour of some entities, and description, where the system findspatterns for presentation to a user in a human-

understandable form [10].A variety of data-mining methods exists which help in achieving the goal ofprediction and description. According to [11], data mining methods commonlyinvolve the following classes of tasks: Inferring rudimentary rules, statistical modelling,constructing decision trees, constructing rules, mining association rules,linear models, instance-based learning, and clustering. For each of these methodsa variety of data mining algorithms exists that can be used. Werefer to [11] for an updated reference on data mining methods and algorithms.By using current modern techniques of big data, AI, and data mining have already been helping the researchers to make an automated system.

1) Machine learning

It is broadly divided into three main categories [12].

- Supervised Learning: This method learns by providing training. The example of input-output pairs is given which is a function learnt by the algorithm.
  Definition: N input-output pairs are provided as training set which is used for approximation taken as y = f(x) where f is referred to as hypothesis and function is used to map input to output space. It can be stated as classification if y is assumed to be a finite set of predefined numbers else regression if y is a continuous number.
- Unsupervised Learning: This method learns itself without any feedback. Unsupervised learning makes use of available lexical resources. In this approach, input given is a set of unknown labels (not known during training) and output comes as classified data. Clustering is a good example of the above technique. Corpus based or dictionary based techniques are used in this type of technique.
- Reinforcement Learning: Reinforcements are considered as a series of feedbacks on which algorithm learns and rewards or punishments are given as the model learns from its decisions.

2) Traditional Supervised Learning Models

Some of the models have been discussed here:

- Naïve Bayes: Classifier is the simplest and the most widely used probabilistic classification algorithm [13. 14]. It is based on Bayes' Theorem.
- Maximum Entropy: is another model which performs probabilistic classification, making use of the exponential model. It is based on the Principle of Maximum Entropy [15] which states that subject to the prior data which has been precisely stated, the probability distribution which describes this data with the current knowledge in the best possible manner is the one with the largest possible entropy value. This technique has been proven to be effective in many NLP classification tasks [16] including sentiment analysis. Max entropy classifier is seen to outperform the Na¨ıve Bayes in many cases [17] (though not always). One major advantage of this classifier is that it makes no conditional independence assumption on the features of the documents to be classified, given a sentiment class. Hence, it is applicable to real-life scenarios, unlike in case of Na¨ıve Bayes.
- Support Vector Machines: Vapnik [18] have proved to be highly effective for the categorization of documents based on similar topics. As opposed to the probabilistic classifiers like the previous two [19] this method aims to find large margin between the different classes. It is a supervised learning model which analyses data and learns patterns which can be used to classify the data. Support vector machines attempt to find a hyperplane (in case of 2-class classification problem) which not only separates data points based on the category they belong to, but also tries to maximize this separation gap between the two classes, i.e., this is a constrained optimization problem. One major advantage of this classifier is that it makes no assumption on the documents to be classified and it endeavors to find the best classification margin for the data at hand instead of relying on probability values. It is one of the widely used machine learning algorithms, which yields very good results for the task of sentiment analysis [17].

III. PROPOSED WORK

Strategic businessIT alignment deals with the application of IT in an appropriate and timely way to meet business goals. A number of empirical studies have been made [20,21,22] that found strategic IS alignment influencing business performance. The purpose of this research is to assess the performance of various data mining techniques when applied to business process models. The methodology consists of three parts:

- An IT infrastructure to facilitate our research, whichincludes a common process model, attributes andprocesses, modeling parameters, model results, etc.
- A model-independent knowledge discovery process is employed to predict the appropriate previously developed model, by using datamining techniques.
- A set of measurements to quantify the similarity of proposed model with previously developed models by different modeling tools, such as naïve bayes, decision tree and neural networks.

A. Architecture:

Process models are represented with the help of dependency graphs [23]. Nodes and edges are used to represent the methods and relationships between them respectively. The arrows representing relations between nodes can be made as predecessor nodes and successor nodes forming hierarchy. The high level abstraction is achieved by presenting a conceptualize view of the business process models (2). These graphs can be extended to allow interorganization process models and can be collaborated with cloud architecture as explained by [24]. The model is divided into three types of blocks as mentioned in Fig.2 :

Primitive, abstract and complex.

The primitive block is the simplest among all which are indivisible as cannot be divided further. All the basic data structures, attributes constitute this block. The two parts are referred as action part and argument part which can be identified as process primitive and product primitive respectively. The former defines the operations and overall flow of the production and the latter describes in the product model.

The abstract block describes the hierarchical relationship between classes and subclasses. It describes all the activities and tasks associated with them.

The complex block present more details of the tasks and activities. They are developed from simpler ones. All the subtasks are detailed out in more detail.After the business process model is developed, we integrate it with techniques of data mining.

Data Mining becomes more and more an integral part of executing a business. The data mining project basically consists of six different phases, according to CRISP-DM. Business Understanding states that the objectives and requirements of the project are thoroughly considered and transformed into data mining problem with respect to business point of view.

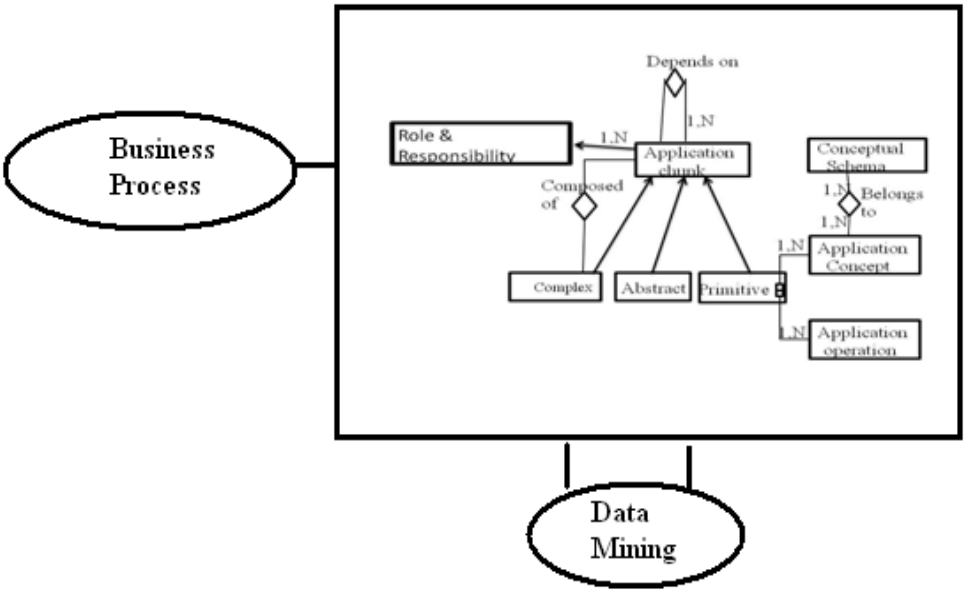

Fig 2. Proposed Architecture

The below diagram explains each term.Fig. 3 shows the conceptual infrastructure that we use to assess the performance of various business process models built via data mining techniques. We built the infrastructure on a data warehouse with tables, views, and micros to facilitate the following model development and assessment processes.

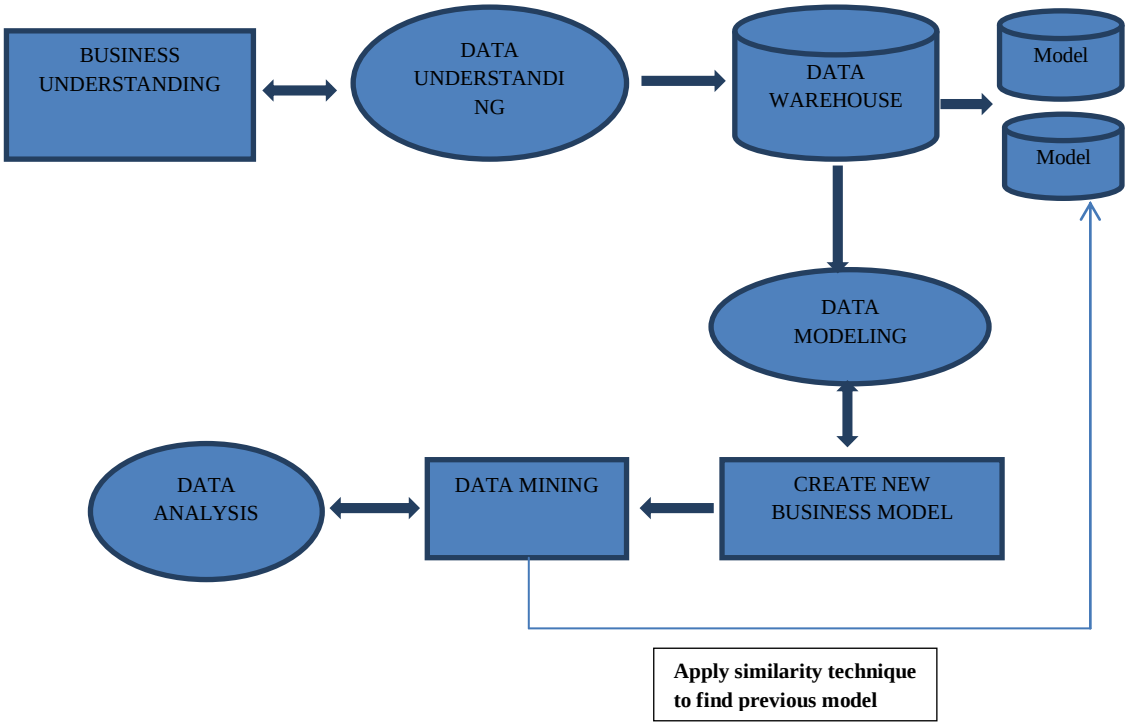

Fig 3. Conceptual infrastructure of model

(1) Data Understanding: Identify data items of interest as Business Process models developed using tool. Extract, transform, and derive variables from identified data items,
The data is prepared by performing cleaning tasks and stored in the data warehouse. The process models are extracted and stored as separate data repositories.
(2) Data Modelling: The Advanced BPMN is developed using tool.
(3) Data Analysis: Various data mining techniques are selected and applied, to choose the appropriate data model stored previously. Modelling techniques are deployed and data mining techniques are evaluated using search strategies.

B. Methodology:

Data mining can be taken as Knowledge discovery from databases. Various data mining techniques explained earlier can be efficiently be applied inorder to choose the appropriate business model.We have employed two techniques, one is using supervised learning technique and other is unsupervised learning technique.

1) One is integrating it with K means clustering:

Cluster Analysis is one of the most important data mining techniques which help the researchers to analyze the data and categorize the attributes of data into various groups. KMeansis one the frequent partitioning algorithm used in clustering [25]. The enhancement of K-means clustering can be done by choosing appropriate initial cluster centers to converge quickly to the local optimum.The steps of conventional K Means algorithm are as follows:

1. Select K points as the initial centroids.
2. Repeat
3. Form k clusters by assigning all points to the closest centroid.
4. Recompute the centroid of each cluster
5. Until centroids do not change.

The potentional function is given by the formula:
Fk means= $\Sigma \Sigma$ $\left| p\text{-}mi \right|$
Where Ci is the i'th cluster, p is a data point, and mi is the
center of the i'th cluster.

But the drawback of the same is that it is sensitive to initial centroids or seeds. If they are chosen wrong, then the results will not be promising. So, to decide the initial centres, we have devised the similarity formula.

- Similarity method:

If two business models have similar features, they usually point to same or similar information needs. The formula for calculating the content based similarity between two models is given as below:

Simkeyfeatures(x, y) = KW (x,y)| |kw (x) $\cup$ kw (y) (eq 1)

where KW(x,y) represents the set of common features in the models x and y, kw(x) and kw(y) are the sets of features in reviews x and y respectively.The threshold parameter used in the above formula is taken as 0.04 by conducting various experiments on different models.if (similarity >= 0.04), the particular model is chosen to be the centroid.

2) Second is integrating it with Naïve Bayes method:

The important features of the model will be identified and supervised learning technique i.e., Naïve bayes is applied to calculate the similarity.

Naïve Bayes: Classifier is the simplest and the most widely used probabilistic classification algorithm. It is based on Bayes' Theorem. It basically calculates the posterior probabilities of events and assigns the label with the maximum posterior probability to the event. A major assumption made by the NB Classifier is that the features are conditionally independent, given the sentiment class of the document [17] which is not true in real-life situations. Furthermore, another problem with this technique is that, if some feature value, which was not encountered in the training data, is seen in the input data, its corresponding probability will be set to 0. Bayes classifier fails in this case. To remove this undesirable effect, smoothing techniques are applied [26].

Both of the supervised and unsupervised method explained above can be used to find the similarity. The efficiency of both the techniques will be compared by search evaluation strategies like precision, recall and f-measure.

IV. ANALYSIS

The quality of the data mining model can be evaluated in terms of few parameters defined in KDD process[27]. The objectives [28] are met for the proposed the data mining model can be evaluated in terms of search startegies like Precision, Recall and F-Measure. . The basis for evaluation is performance metrics such as Precision, Recall and F-measure which are used for calculating accuracy in our work.

Precision: It is defined as the ratio of the correctly identified model over all the models stored. Mathematically, the Precision is given by: P = OM / (OM+WOM), where OM is Relevant Models identified and WOP is the number of Irrelevant Models identified. It is usually expressed as a percentage.

Recall: It is defined as the ratio of the fraction of correctly identified model by the previously models stored over all the models as given by the experts. Mathematically, the Recall is given by: R = OM/ (OM + NOM), where OM are Relevant models identified and NOM Relevant models not identified. It is usually expressed as a percentage.

F-Measure: It is the combination of recall and precision. Thus, F-measure combines these two values. F = 2PR/ (P + R), where Precision P and Recall R are equally weighted.

Apart from these evaluation criteria's, other important factors like the time spent forthe process, the resources that were used, etc., can be included in the evaluation. In the evaluation phase of the data mining process, the model as well as the wayit was constructed is evaluated according to the business objectives.

We have taken few previously stored models with the particular business objectives. Therefore, by conducting experiments, it was found that the results with the K means is much high in terms of precision and recall. In total, a sample of 20-25 different models have been gathered and analyzed shown in Table 1.

Table 1. Results

No. Of models	Precision (%)	Recall (%)	F-Measure (%)
0-5	81.61	82.66	82.13
5-10	85.7	85.86	85.77
10-15	87.82	89.9	88.84
15-20	91.27	93.73	92.48
20-25	92.1	94.2	93.64

The accuracy of the proposed data mining model using unsupervised learning techniques is better than supervised learning technique is shown in Fig 4.

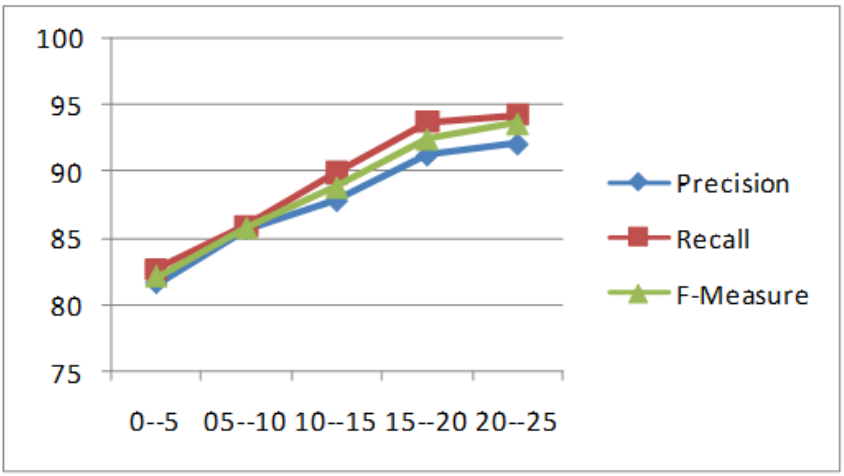

Fig 4. Analysis

V. CONCLUSIONS

BPM and Data mining are complex areas and require a lot of reengineering. But in many practical applications, they are showing remarkable improvements. We have developed guidelines for developing the BPMN model and also explored the data mining techniques to find the similarity factor between the proposed model and previously built models. In this paper, we have dealt with two different techniques of data mining. The supervised learning algorithm and unsupervised learning techniquehas been proposed to find the most similar models previously developed with the proposed ABPMN model. Integrating data mining into business processes has proved beneficial in many systems of Ecommerce as well. As future work we foresee improvement on the concepts and techniques foran easy deployment of flexible data mining solutions into business processes in the context of modern BPM frameworks based on BPEL and BPMN, based on a uniform concept for both modelling and technical integration. The propose model will be simulated later under different scenarios of cloud manufacturing with data mining to clearly understand the effective integration between data mining services that support an easy and business processes.

REFERENCES

[1] Wegener, D., &Rüping, S. (2010, May). On Integrating Data Mining into Business Processes.In BIS (pp. 183-194).
[2] Rozinat, A., & Aalst, W. M. P. (2006). Decision mining in business processes. Beta, Research School for Operations Management and Logistics.
[3] Koskela, M., &Haajanen, J. (2007).Business Process Modeling and Execution. Tools and technologies report for SOAMeS project.
[4] Jordan, D., Evdemon, J.: Web Services Business Process Execution Language Version 2.0. Technical report, OASIS Standard (2007)
[5] White, S. A., Miers, D.: BPMN Modeling and Reference Guide Understanding and Using BPMN. Future Strategies Inc., Lighthouse Pt, FL (2008)
[6] Peisl, R.: The Process Architect: The Smart Role in Business Process Management. IBM RedPaper (2009)
[7] Weske, M. (2012). Business process management architectures.In Business Process Management (pp. 333-371).Springer Berlin Heidelberg.

[8] Weske, M. (2010). Business process management: concepts, languages, architectures. Springer Publishing Company, Incorporated.
[9] Laswad, A. T., Ghazy, M. M., &Elsayed, A. E. Development of a Business Process Model Notations (BMPN) for a Cloud Manufacturing Architecture.
[10] Fayyad, U., Piatetsky-Shapiro G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. AI Magazine 17, pp. 37{54 (1996)
[11] Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco (2005)
[12] Russell, S., Norvig, P., & Intelligence, A. (1995). A modern approach. Artificial Intelligence.Prentice-Hall, Egnlewood Cliffs, 25, 27.
[13] Russell, S. J., & Stuart, J. (2003). Norvig. Artificial Intelligence: A Modern Approach, 111-114.
[14] McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, pp. 41-48).
[15] Jaynes, E. T. (1957). Information theory and statistical mechanics. Physical review, 106(4), 620.
[16] Berger, A. L., Pietra, V. J. D., &Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. Computational linguistics, 22(1), 39-71.
[17] Pang, B., Lee, L., &Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86).Association for Computational Linguistics.
[18] Vapnik, V., & Cortes, C. (1995). Support Vector Networks, machine learning 20, 273-297.
[19] Joachims, T., Nedellec, C., &Rouveirol, C. (1998, January). Text categorization with support vector machines: learning with many relevant. In 10th European Conference on Machine Learning.
[20] Chan Y.E., Sabherwal R., and Thatcher J.B., Antecedents and Outcomes of Strategic IS Alignment: An Empirical Investigation, IEEETEM, 53, 1, 27-47, 2006
[21] Malik K., and Goyal D.P., IS Alignment and IS Effectiveness: Experiences from Indian Industry, IEEE, 96-100, 2003
[22] Henningsson S., Svensson C., and Vallen L., Mastering the Integration Chaos Following Frequent M&As: IS Integration with SOA Technology, Proc. 40th HICSS, 2007
[23] Koehler J, Hauser R., Kuster J, Ryndina K, Vanhatalo J, and Wahler M, The Role of Visual Modeling and Model Transformations in Business-driven Development, Proc. 5th Inti Workshop on Graph Transformations and Visual Modeling Techniques, 1-12,2006
[24] Prakash, N., &Chaturvedi, A. K. (2010, May). Representing analysis models for alignment.In Research Challenges in Information Science (RCIS), 2010 Fourth International Conference on (pp. 409-414).IEEE.
[25] Bhatia, S. (2014, April). New improved technique for initial cluster centers of K means clustering using Genetic Algorithm. In Convergence of Technology (I2CT), 2014 International Conference for (pp. 1-4). IEEE.
[26] Chen, S. F., & Rosenfeld, R. (2000). A survey of smoothing techniques for ME models. IEEE transactions on Speech and Audio Processing, 8(1), 37-50. Jaynes, E. T. (1957). Information theory and statistical mechanics. Physical review, 106(4), 620.
[27] Fayyad, U., Piatetsky-Shapiro G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. AI Magazine 17, pp. 37{54 (1996)
[28] Hornick, M.F., Marcad, E., Venkayala, S.: Java Data Mining: Strategy, Standard, and Practice. Morgan Kaufmann, San Francisco (2006)

