Journal of Theoretical and Applied Information Technology
31st July 2017. Vol.95. No 14
© 2005 – ongoing JATIT & LLS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ALTERNATIVE MODEL BASE AS AN ENABLER FOR SUCCESS OF BUSINESS INTELLIGENCE-BASED COMPANIES USING C4.5

KISWORO
Universitas Teknokrat Indonesia, Faculty of Technic and Computer Science, Indonesia
E-mail: kisworo@teknokrat.ac.id

ABSTRACT

Business intelligence has been base for new form of business in the world. Companies have respectfully adopted the base for gaining success. Internet is used for reaching the goal. Information and users becomes core assets for facing the future. But so many complex factors have been making up every detail efforts. It is very interesting to be researched under conscience and smartness. Business intelligence concepts should be observed and causes should be found. From computer science side, especially computation with its effects to decision making of management level, this research has found a small side relates to alternative model base as enabler. From ocean names computation references, C4.5 appears and looks very incredible to be entered. A decision then comes to use by the reasons that the algorithm has been top ten in the computation world. It is used widely by academic environment and practitioners, tested and accepted. It will be promising for the future. Therefore, this research has: (a) motivation to encourage seeking answer against hope for success in Business Intelligence-based business, (b) focus on offering an alternative model base possibility for companies to develop information, (c) research object: companies spread on the internet, and (d) significance on contributing modeling result in form of decision tree via C4.5.

Keywords: *Business Intelligence, Alternative Model Base, C4.5, Companies.*

1. INTRODUCTION

Business Intelligence (BI)-based companies need to be presented here because of their core assets, i.e. information[3, p.23]—please no *Information Anxiety*[34]—are very promising relates to decision making[26, Ch.2] in the latest era, information age[1]. It is not "*The Great Promise*"[8] because "*being*" (v.s. "*having*") *an sich* is always increasing curiosities on process learning to people-focused knowledge management[32] of BI-based users, i.e. those who are capable of discovering knowledge in data[17]. There is knowledge mining[14] inside, at where knowledge management looks fulfilling exhibit: *data-information-metadata-knowledge-understanding* [4, p.11]. (See, for example, caused by successful BI[12] in problems complexity of information environment[3] and successful decision making[11], Zadeh had ever purposed "*concept*" (*text* and *images*) for search engine[37] with his evolutionary fuzzy logic[36]). The exhibit ranges explicit and tacit knowledge in various implementations of decision support systems for BI[26]. Works in team/group of BI completed by roadmap[21] and human capital in scale of enterprise knowledge with adequate infrastructure[18] now have orientation to knowledge-based organizations[7]. Knowledge management as a layered multi-disciplinary pursuit[27] saves consequences that BI-based companies should be innovative[29, Ch.5], competitive[20], and adaptive[19][29, Ch.6]. Internet enables the situation[9].

BI as base of companies business activities is one of decision process making components[30]. Related to other components, i.e. decision support system and data warehousing, BI is a set of mathematic model and analytic methodology exploit given data for information and knowledge used by complex decision-making process[35]. In business analysis scale, IBM caught the phenomena as the new promise of business intelligence[13]. It is said: "*A lot has changed since business intelligence first came on the scene promising to help IT deliver information to the business. Information assets have grown exponentially, business users have become far more proficient with technology, and the pace of business has* *accelerated. And so business intelligence must evolve too.*" In other place, [31] said: "*A growing number of companies are becoming BI-based. For these firms, business intelligence is not just nice to have; rather, it is a necessity for competing in the marketplace. These firms literally cannot survive without BI ...*"

Therefore, this research has: (a) motivation to encourage seeking answer against hope for success in BI-based business with users as other core assets[4], (b) focus on offering an alternative model base possibility for companies to develop information, (c) research object: companies spread on the internet, and (d) significance on contributing modeling result in form of decision tree via C4.5[24]. Related to the algorithm, [16] said: *"C4.5 ... is a suite of algorithms for classification problems in machine learning and data mining. ... In addition to inducing trees, C4.5 can also restate its trees in comprehensible rule form. Further, the rule post pruning operations supported by C4.5 typically result in classifiers that cannot quite be restated as a decision tree."*

2. RESEARCH METHODOLOGY

The research model uses the following CRISP-DM methodology:

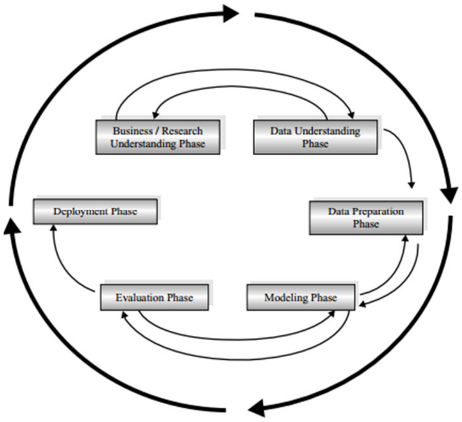


Figure 1: CRISP-DM Methodology

Figure 1**:** CRISP-DM Methodology is adopted from [17, Ch.1:p.6] without a change of naming for the figure scheme, except a change of naming for the name of the figure. The name of the figure is CRISP-DM, stands for Cross-Industry Standard Process for Data Mining. CRISP-DM stands for Cross-Industry Standard Process for Data Mining. Analysts developed the methodology in 1996. It is adjusted with a given data mining project has a life cycle consists of six phases. All of the phases are adaptive.

3. RESEARCH BRIDGE

Modeling of this research uses the following research bridge:

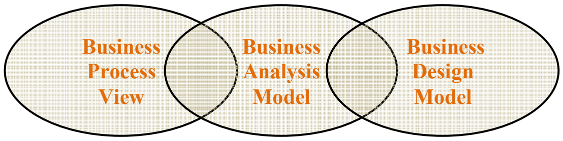


Figure 2: Research Bridge

Figure 2: Research Bridge above is adopted from Pressman[22, Ch.6] with changes in naming. The name of the above figure: Research Bridge. Its basic idea is similar to Pressman's, i.e. the middle oval is a bridge for the other both within intersection. The next explanation is as the below:

a. Business Process View (BPV) is placed before BAM, meaning: worked first; and its position is lower than BAM, meaning: when working activities, BVP works also a part of BAM activities while looking BAM above. Core of BPV activity is to catching *as-is* business process existence. It is done in requirements engineering scope. The factor is very important because of customers' evolutionary needs. Theoretically, there are seven phases could be worked through the requirements engineering, i.e. *inception*, *elicitation*, *elaboration*, *negotiation*, *specification*, *validation*, and *management*. Inside all of the seven phases, refinement always should be flexible.

b. Business Analysis Model (BAM) is placed in the middle, between BPV and BDM, meaning: when working activities, BAM works BPV and BDM activities while looking the both from the height; the position depicts activity core of modeling process. It is done because of business modeling process gets a position core for analysis process. For this research, a work set of C4.5 is used after requirements identification process.

c. Business Design Model (BDM) is placed after BAM, meaning: worked after BAM; and its position lower than BAM, meaning: when working activities, BDM works also a part of BAM activities while looking BAM above. It happens because business modeling process gets its design core here. For this research, implementation could be done by companies get its milestone, i.e. an offered model base, an alternative model base resulted by this research.

4. RESEARCH FRAMEWORK

Modeling of this research uses the following research framework:

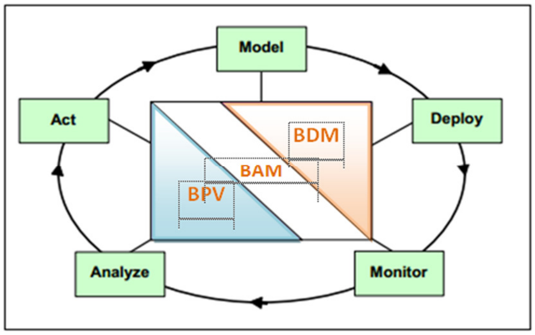


Figure 3: Research Framework

Figure 3: Research Framework above is adopted from Ballard, *et al*[2, p.6-7] with changes in naming. The name of the above figure: Research Framework. The name refers to integrated process cycle. For this research, the activities used are Analysis, Act, and Model. Analysis and Act are done in BPV, while Model is done in BDM. The three is not connected vividly with BAM, but their activities are related roughly with and never disconnected by BAM. BAM becomes central point (core) of the other both, to which their cycle are based on the research bridge.

5. DATA MINING (DM)

5.1 DM in KDP

Solution for the above question at point 2 is machine learning technology [5,p.2]. Machine learning is not intended to mechanical machine here but to technique of finding and describing structural patterns in data[33, Ch.1]. The important part of the technology is DM. DM is process of discovering patterns in data; such patterns are structural patterns[33, Ch.1]. Process used by DM is knowledge discovery (KD), i.e. a walk through operations and nuances of various algorithms like C4.5 algorithm using datasets so that it gets true appreciation of what is really going on the inside of the algorithm[2, p.xii). Through the process, DM has just one place in the following KDP:

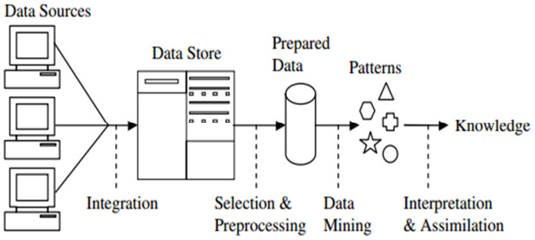


Figure 4: The Knowledge Discovery Process (KDP)

As central to knowledge discovery, DM does not stand alone. Related to the near neighbor of DM, there is a phase before, i.e. prepared data, and there is a phase after, i.e. interpretation.is proceeded by patterns, resulted by DM algorithm[5, p.3].

5.2 DM Process

The existence of DM in context of the KDP should be widened trough explaining the DM process. The core of the DM process is as the following:

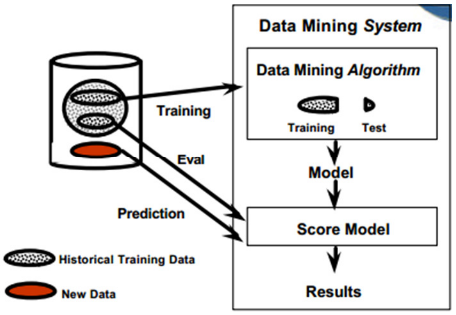


Figure 5: Data Mining Process[29, slide. 5]

Some important elements are involved inside the DM process. Inputs could be processed through training, evaluation, and prediction. Then DM system works within itself through DM algorithm, model, score model, and results. DM algorithm and score model accept inputs and process them. This research is for alternative model base, but score model and result are involved inside.

5.3 DM Task: Classification

Tasks had by DM are description, estimation, prediction, classification, clustering and association[17, p.11]. This research will use classification. In classification, there is target categorical (nominal and ordinal) variables that could be partitioned into classes or categories. It is examines a large set of records. Each record contains information in the target variable as well as a set of input or predictor variables. The classification is under an algorithm. This research uses C4.5 algorithm.

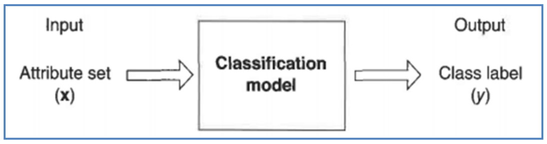
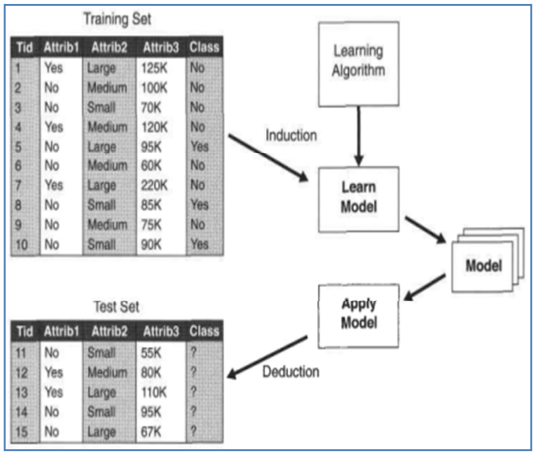


Figure 6: Input Attribute Set x to Its Class y[28, p.146]

5.4 Solving Classification Problem: General Approach

Classifier is a classification technique, a semantic approach to building classification models from an input data set. Each technique employs a learning program algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. The model generated by a learning algorithm should fit the input data well and correctly predict the class labels of records it has never seen before. Key objective of the leaning algorithm is to build models with good generalization capability.



Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Test Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Figure 7: General approach for building a classification model[28, p.148]

5.5 Classification: DT

DT stands for Decision Tree, a classification method of DM, an attractive part of DM, constructed by DM process, and used as reasoning procedure. It is a collection of decision nodes, connected by branches, extending downward from root node until leaf node. The DT begins at the root node at the top and terminates at the leaf node at the end. The root node does not have an input but has zero or more output(s). Reversely, the convention is for the leaf node also. Among of them called as internal node(s) is for testing every feature value at every branch before and after the given node[10, Ch.4] [28, p.150-151].

The formed DT is not always binary tree. If all features in data set using two kinds of categorical values, form of the tree is the binary tree. If all features in data set using more two kinds of categorical values, or using numerical ones, form of the tree is not binary tree as usual. The DT is applied in many real world[15, p.170].

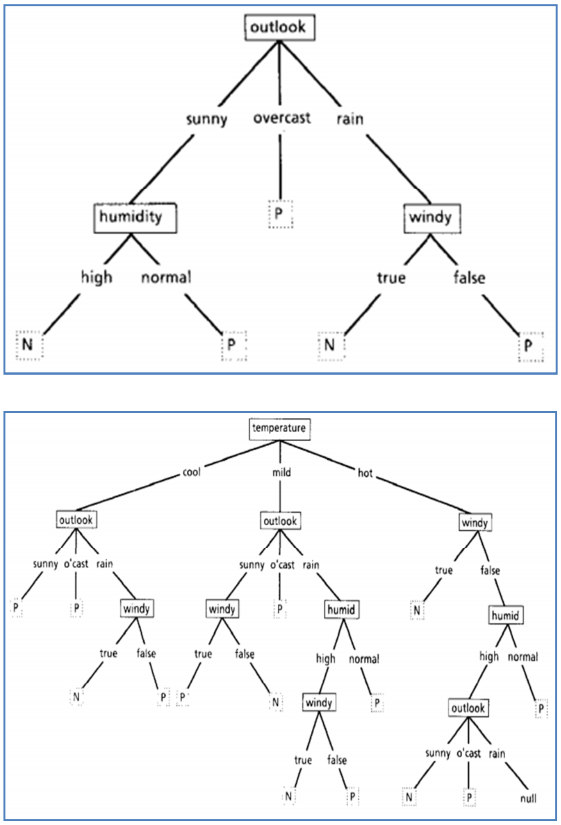
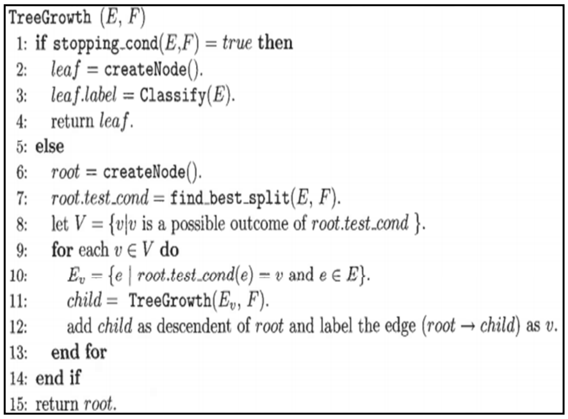


Figure 8: Simple and Complex Decision Tree[23, p. 87-88]

6. C4.5 ALGORITHM

The presence of C4.5 is very important to increase performance limitations belonged to its previous version, ID3. C4.5 could induct DT with categorical and numerical type features, could prune DT, and could derive rule set. DT induction is about how to state testing requirements to a node. There are three features for the requirements: (1) binary feature with only two branches to be opted, (2) categorical (numerical or ordinal) type feature possibly with some difference values, that commonly has two splitting kinds, i.e. binary splitting (with 2^{k-1}-1 splitting combination(s)) and multi splitting, and (3) numerical type feature with root and internal node testing requirements stated by comparison testing A < *v* or A ≥ *v* with resulting in binary splitting, or for multi splitting with resulting range $v_i \leq A \leq v_{i+1}$ where i = 1, 2, …, k. For binary splitting case, algorithm would investigate all position possibilities of split *v* and chooses the best for position *v*. For multi splitting, algorithm should investigate all range possibilities of continual values.

C4.5 algorithm used here is as the below:



```
TreeGrowth (E, F)
if stopping_cond(E,F) = true then
    leaf = createNode().
    leaf.label = Classify(E).
    return leaf.
else
    root = createNode().
    root.test_cond = find_best_split(E, F).
    let V = {v|v is a possible outcome of root.test_cond }.
    for each $v \in V$ do
        $E_v = \{e \mid$ root.test_cond(e) $= v$ and $e \in E\}$.
        child = TreeGrowth($E_v$, F).
        add child as descendent of root and label the edge (root → child) as v.
    end for
end if
return root.
```

Figure 9: C4.5 Algorithm[28, p.164]

The original thought released by Quinlan [24, p. 17-18] could help catch recognition against the algorithm.

7. CONSTRUCTING DT VIA C4.5 STEPS: DT INDUCTION

Based on some references: [15, Ch.6][28, Ch.4][24, Ch.2][17, Ch.6][6, Ch.5], it is said that DT induction has steps that respectively should work in process of constructing DT. The steps could appears because of the C4.5 algorithm used(point 7). The meant steps are: entropy, gain, ratio gain, and IF-THEN rule. The peak of the steps is the ratio gain, but the involved IF-THEN rule has useful function on implementing the constructed tree via programming languages as Matlab.

a. Entropy
Entropy is used for determining node that would be splitter for next training data. The higher entropy value would increase classification potency. If entropy for the node has value 0, it means all vector data exists in the same class label and the node should be leaf with decision inside (class label). If one of elements ω_i has total 0, the entropy would be value 0 also. If all elements ω_i has the same total in proportion, the entropy would be value 1. The entropy formula is given by:

$$E(s) = -\sum_{i=1}^{m} p(\omega_i \mid s) \log_2 p(\omega_i \mid s) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where $p(\omega_i \mid s)$ is #*i* class proportion in all training data processed in node *s*. $p(\omega_i \mid s)$ is from total of all data rows with class label *i* divided by row total of all data. *m* is total of different value inside data.

b. Gain
Gain is used for investigating rightly feature option as splitter to the node. Gain is a #*j* feature counted using the below equation:

$$G(s,j) = E(s) - \sum_{i=1}^{m} p(v_i \mid s) \text{ x } E(s_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

where $p(v_i \mid s)$ is portion of value *v* that appears in class inside the node. $E(s_i)$ is entropy competition of value *v* of #*j* class in the #j node. *n* is total of different value inside data.

c. Ratio Gain
With

$$SplitInfo(s,j) = -\sum_{i=1}^{k} \log_2 p(v_i \mid s) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

where k states splitter total, ratio gain as criteria could be used for opting feature as splitter inside C4.5 algorithm. The ratio gain formula is given by:

$$RatioGain(s,j) = \frac{G(s,j)}{SplitInfo(s,j)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

d. IF-THEN Rule
Process a to c described above eventually bears a tree. That's all process of the DT induction. The tree is the DT that could be used for making decision. Then, after, the meant DT could be formed in the way of IF-THEN rule. The way should work with tracing

forward a path. The tracing for the path should be started from the root node followed by the next nodes till the leaf node since from the leftmost path to the rightmost path in the way of one per one path.

8. A RESEARCH TASK: MODEL WITH DYNAMIC CONDITIONS AND REQUIREMENTS[25]

The first, if given by the below data set condition, where a prediction should satisfy the case, what should see is features used such as f1, f2, f3. Then see also data with their types and values inside as f1 = {a1, a2,a3}, f2 = {b1, b2,b3}, f3 = {c1, c2,c3}. The data set is really not flexible, not dynamic or static. It is limited by the variables of every feature and data value. For building flexible-dynamic model, features and data values presented by the following data set table are needed:

Table 1: Dynamic changes of data set

f1	f2	f3	…	fn
a1	b1	c1	…	…
a2	b2	c2	…	…
a3	b3	c3	…	…
…	…	…	…	…
an	bn	cn	…	…

The dots depict the meant flexibility. Requirements come from users/customers who always experience growth. There are possibilities for new features and data values that should be accommodated caused by the coming changes.

The second, entropy should be flexible with the condition. Inside the opted algorithm, C4.5, the flexibility on assigning every input relates to the entropy formula that should be accommodative. Dynamic model should be capable on satisfying the dynamic requirements (come also from users/customers).

The third, if both previous numbers have been done rightly, it gains the mentioned principle and condition that would be affected by the work flow. Ratio gain would find out feature splitters without questionable enrichments.

9. CONCLUSION

Alternative model base offered here really does not only represent visible things. If should be, this research would not fully catch the meant target. Success of BI-based companies surely needs the more. C4.5 keeps abundant secrets that should be researched although C5 has appeared. The world positively appreciates so much with the C4.5 and its generations before and after. The reality would be the promising future task.

REFERENCES:

[1] Alberts, D.S. & Papp, D.S, *The Information Age: An Anthology on Its Impacts and Consequences*, Washington DC: CCRP Publication Series (US), 1997.

[2] Ballard, C., etc, *Business Performance Management Meets Business Intelligence*. available on line at http:/www.ibm.com/redbooks, 2006.

[3] Ballard, C., et.al, *Dimensional Modeling: In a Business Intelligence Environment*. IBM Corporation. Redbooks (US), 2006.

[4] Bergeron, B, *Essentials of Knowledge Management*, New Jersey: John Wiley and Sons, Inc (US), 2003.

[5] Bramer, M, *Principles of Data Mining.* London: Springer-Verlag (UK), 2007.

[6] Chakrabarti, S., et.al,. *Data Mining: Know It All*. Elsevier. Inc.(US), 2009

[7] El Sheikh, A.A.R. & Alnoukari, M, *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations—Cross Disciplinary Applications.* Business Science Reference (US), 2012.

[8] Fromm, E, To *Have or To Be*. New York: Harper and Row (US), 1976.

[9] Giovinazzo, W, *Internet-Enabled Business Intelligence*. Pearson Education, Inc (US), 2002.

[10] Gorunescu, F, *Data Mining: Concepts, Models and Techniques*, Berlin-Heidelberg: Springer-Verlag (Germany), 2011.

[11] Grunig, R. and Kuhn, R, S*uccessful Decision Making: A Systematic Approach to Complex Problems.* Berlin-Heidelberg: Springer-Verlag (Germany), 2005.

[12] Howson, C, *Successful Business Intelligence: Secrets to Making BI a Killer App.* New York: McGraw-Hill (US), 2008.

[13] IBM Software Group, *The New Promise Business Intelligence*. New York: IBM Corporation. White Paper (US), 2010.

[14] Jie, S., et.a., "Knowledge Mining for Web Business Intelligence Platform and Its Sequence Knowledge Model". *International Conference on Computational Intelligence and Security Workshop*, 2007. Available [Online] at en.bookfi.net, Jan. 13rd 2017.

[15] Kantardzic, M, *Data Mining: Concepts, Models, Models, and Algorithms,* IEEE Press, New Jersey: John Wiley and Sons, Inc (US), 2011.

[16] Wu, X and Kumar, V, *The Top Ten Algorithms in Data Mining*. London: Chapman and Hall (UK), 2009.

[17] Larose, D.T, *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: John Wiley and Sons, Inc (US), 2005.

[18] Maier, R., Hadrich, T., and Peinl, R, *Enterprise Knowledge Infrastructure*. Berlin-Heidelberg: Springer-Verlag (Germany), 2005.

[19] Michalewicz, Z, et.al., *Adaptive Business Intelligence*. Berlin-Heidelberg: Springer-Verlag (Germany), 2007.

[20] Miller, G.J., Brautigam, D., and Gerlach, S.V., *Business Intelligence Competency Center: A Team Approach To Maximizing Competitive Advantage*. New Jersey: John Wiley and Sons, Inc (US), 2006.

[21] Moss, B.L.T. and Atre, S, *Business Intelligence Roadmap: The Complete Project Lifecycle and for Decision-Support Applications.* Boston: Pearson Education, Inc (US), 2003.

[22] Pressman, R.S, *Software Engineering—A Practitioner's Approach*. New York: McGraw-Hill. 7th Edition (US), 2010.

[23] Quinlan, J.R, "Induction of Decision Trees", *Journal Machine Learning*, Vol.1, No.1 1986, p. 81-106.

[24] Quinlan, J.R, *C4.5: Programs for Machine Learning*, New York: Morgan Kauffman (US), 1993.

[25] Quinlan, J.R, "Improved Use of Continuous Attributes In C.4", *Journal of Artificial Intelligence Research,* Vol. 4 , 1996, p.77-90.

[26] Sauter, V.L, *Decision Support Systems for Business Intelligence*. New Jersey: John Wiley and Sons, Inc. 2nd Edition (US), 2010.

[27] Schwartz, D.G, *Encyclopedia Knowledge Management*. Idea Group Reference (US), 2006.

[28] Tan, P-N., Steinbach, M., and Kumar. V, *Introduction to Data Mining.* USA, Boston: Pearson Education, Inc (US), 2006.

[29] Thearling, K. --------. *An Introduction to Data Mining*. Available [Online] at en.bookfi.net 18 Jan. 2017. In form of slide.

[30] Vercellis, C, *Business Intelligence: Data Mining and Optimization for Decision Making*. New Jersey: John Wiley and Sons, Inc. (US), 2009

[31] Wixon, BH and Watson, J.H, "The BI-based Organizations". *International Journal of BI Research*, Vol. 1, No. 1, 2010, p.13-28.

[32] Wiig, K., *People-Focused Knowledge Management: How Effective Decision Making Leads to Corporate Success.* Elsevier. Inc (US), 2004.

[33] Witten, I.H., Frank, E., and Hall, M.A, *Data Mining: Practical Machine Learning Tools and Techniques*, Elsevier. Inc. 3rd Edition (US), 2011.

[34] Wurman, R.S, *Information Anxiety*. Indianapolis: Que Pusblishing (US), 2000.

[35] Zadeh, L.A, "Outline of A New Approach to the Analysis of Complex Systems and Decision Processes". *IEEE*, Vol. 3, No.1, 1973.

[36] Zadeh, L.A, "The Birth and Evolution of Fuzzy Logic" *International Journal General Systems*, Gordon and Breach Science Publishes (UK), Vol. 17, 1990, p. 95-105.

[37] Zadeh, L.A, "Web Intelligence, World Knowledge and Fuzzy Logic -- The Concept of Web IQ (WIQ)", In M.Gh, Negotia, et.al. (eds), *Knowledge-Based Intelligent Information and Engineering Systems. KES 2004. Lecture Notes in Computer Science* , Springer (Berlin), V, LNAI. 3213, 2004, p.1-5.