OPEN ACCESS Journal of Artificial Intelligence

ISSN 1994-5450
DOI: 10.3923/jai.2017.

Research Article
Data Mining Approach for Detecting Key Performance Indicators

[1]Nehaya Sultan, [2]Ayman Khedr, [3]Amira Idrees and [1]Sherif Kholeif

[1]Faculty of Computers and Information, Helwan University, Helwan, Egypt
[2]Faculty of Computers and Information Technology, Future University, Cairo, Egypt
[3]Faculty of Computers and Information, Fayoum University, Faiyum, Egypt

Abstract

Background and Objective: Key performances indicators (KPIs) are an integral part of business intelligence systems as the choice of KPIs are critical to success. This study aimed to propose a solution to detect KPIs from historical organizational data using data mining algorithms and analyzes the relation between factors that will affect the performance to help organizations execute their business strategy. This approach does not involve domain experts to identify or validate KPIs. **Materials and Methods:** Information gain algorithm implemented with Weka (InfoGainAttributeEval) used for feature selection to rank the attributes that affect the performance. Moreover, an improved FP-growth algorithm was used to find the correlation between attributes. **Results:** The KPIs detection model was tested using 6 years of banking data. To detect the non-performing loans KPI the model indicates strong correlation between non-performing loans and attributes such as accounts with low salaries, young clients and accounts with a monthly issuance of statements. **Conclusion:** The proposed KPI detection approach can make the process of selecting KPIs more efficient; it can be used as a method to determine the most appropriate KPIs. This model will enable decision makers to make timely and appropriate strategic decisions.

Key words: Business intelligence, key performance indicators, data mining, feature selection, information gain, FP-growth

Received: Accepted: Published:

Citation: Nehaya Sultan, Ayman Khedr, Amira Idrees and Sherif Kholeif, 2017. Data mining approach for detecting key performance indicators. J. Artif. Intel., CC: CC-CC.

Corresponding Author: Nehaya Sultan, Faculty of Computers and Information, Helwan University, Helwan, Egypt Tel: +201127649444/+96597781138

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Business intelligence systems combine analytical tools with operational data to present complicated and competitive information to decision makers[1]. Key performance indicators (KPIs) are an integral part of business intelligence systems as the choice of KPI is critical to success[2]. Organizations usually identify KPIs based on reference industry KPIs or a predefined list that may not be relevant to the specific situation and organization and may waste an organization's time and resources in tracking incorrect indicators[3]. Moreover, the traditional process of experts selecting the appropriate KPIs requires considerable experience in the domain. This study proposed a solution to detect KPIs based on historical organizational data using Data Mining (DM) algorithms and analyzes the relation between factors that affect the performance, to help organizations execute their business strategy. The advantage of this method is selecting the right KPIs without prior experience in the domain.

Many studies have focused on identifying KPIs in various sectors[4,5] using traditional means but few have focused on their identification and selection using data mining. Peng *et al.*[6] used a semi-automatic system with iterative learning processes for analyzing operational metrics, filtering KPIs and discovering leading indicators. They identified KPIs with the help of domain experts and used dimensionality reduction techniques to filter them. To discover leading indicators, they explored correlations among reduced indicators by considering time shifts. This was an iterative process that continuously discovered leading indicators along with business change. However, in this process, domain experts are needed to identify and define KPIs. Ishak and Sahak[7] used standard ISO KPIs for library systems but the predefined list may not be suitable for small libraries. Granberg and Munoz[8] used a questionnaire to collect key information to select KPIs but this process is organization specific. Ross and Ingo[9] introduced a machine learning application to select the most important KPIs for call center agents at the initial stage of customer inquiry. These KPIs were selected based on a large database of available KPIs and business performance results. The researchers used support vector machine for classification, which does not work well with larger data sets and there was a limitation of noise when applying the method to real data. Stefanovic[10] introduced two different approaches for building KPI prediction models-the first uses online analytical processing (OLAP) DM dimensions where the results of predictive calculations are saved in a new dimension in the OLAP cube and the second uses prediction tables. In this approach, DM predictions are executed within the Extract Transform Load (ETL) process[10], this approach is more flexible because more tables can be added to the Data Warehouse (DW). As the model is defined outside the cube, it can be changed or replaced without altering the cube itself. However, the model does not include a specific predication method. This approach predicts the value of the KPI but does not detect the KPI itself[10].

It is suggested that few design methods are available to select and detect KPIs associated with business goals that do not need a predefined KPI list or prior experience in the domain[4-10]. The DM techniques are often used to discover trends, patterns and associations and thus this study proposes a solution to detect KPIs based on historical data by using association rules to discover the relations between attributes. Weka's[11] Information Gain (IG) algorithm used to select the top factors affecting performance. An improved Frequency-Pattern (FP)-growth algorithm was then applied to discover the correlation between attributes. The detected KPIs are correlated with the organization goals to identify the factors that influence the performance of each indicator. This study contributed to the field of study on the detection of KPIs as well as feature selection and association rules. It also presents a new way to detect KPIs without the need for prior domain experience or a predefined list, unlike other studies[6,7]. This proposed framework could be used as a general method to determine the most appropriate KPIs, which help organizations to measure the progress towards their business goals.

MATERIALS AND METHODS

Detection models based on historical data were described in this study.

Feature selection: Sutha and Tamilselvi[12] stated that feature selection was one of the most important pre-processing steps used to improve mining performance by reducing data dimensionality before applying techniques such as association rules, classification and clustering[12]. The filter approach of feature selection was used, which includes a pre-processing step independent of the induction algorithm. Weka's IG algorithm (InfoGainAttributeEval) was used to rank the most critical factors that influence the business goals.

Association rules: Association rules mining-one of the most important topics in DM research-aims to extract interesting associations, frequent patterns, correlations or casual structures in transaction databases. Han *et al.*[13] developed the

FP-growth algorithm to discover frequent item sets without candidate generation by constructing a prefix tree (FP tree) to compress the database and then used divide-and-conquer to divide the tree into smaller trees (known as conditional trees) to mine the frequent item sets separately. Recent studies[14,15] showed that FP-growth is one of the most effective frequency pattern algorithms. However, it recursively constructs conditional trees, requiring more memory and time for mining. This study presents an improved FP-growth algorithm that combines FP-growth with Compact Pattern tree (CP-tree)[16,17]. This improved algorithm supports interactive and incremental mining and scans the database just once.

Algorithm 1 (Improved FP-growth):

Input: A transaction database DB and support threshold minimum support (minsup)

Output: CP-tree

Method: CP-tree is constructed as per the following steps:

(1) Scan the transaction database DB once

(2) Insert item into the tree according to the order of the item's appearance

(3) Collect the set of frequent items (F) and their supports

(4) Rearrange the list in descending order of frequency and keep only frequent items
(whose frequency count is greater than minsup)

(5) If the path is not sorted according to the new list order, it is removed from the tree, non-frequent items are deleted and items are sorted according to the new list order into a temporary array and again inserted into the tree

Proposed framework: The key concept behind detecting KPIs using DM algorithms is discovering the correlations between attributes from historical data to identify the most important factors that influence performance. The eight phases of proposed framework is illustrated in Fig. 1.

After setting the business goals, the data is made consistent. The database source is then built with appropriate relationships using SQL server management studio. In order to detect KPIs using association rules, the attributes are discretized and interval boundaries are made consistent. Data is also aggregated to reduce the number of rows to be queried. Aggregation involves rolling up huge amounts of data into higher levels of dimension hierarchies. Microsoft SQL Server Analysis Services (SSAS) is used to build the DW as it provides a complete platform for data warehousing and business intelligence with high efficiency and simplicity[18]. The DW design proceeds in three steps: Designing the DW, the ETL and the OLAP cubes. The original DW consists of two OLAP cubes.

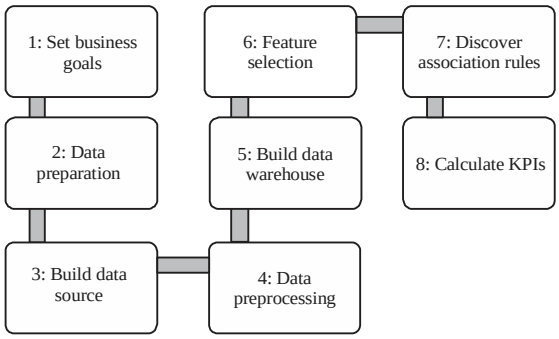

Fig. 1: Proposed framework to detect KPIs using DM algorithms

Table 1: Correlation coefficients matrix

Variable vs. variable	R
Non-performing loan vs. payments	0.18244
Non-performing loan vs. amount	0.16753
Age vs. non-performing loan	0.75203
Non-performing loan vs. duration	0.02582
Non-performing loan vs. salary	0.52448
Non-performing loan vs. frequency	0.82966
Gender vs. non-performing loan	-0.00722

Feature selection was used to select a subset of variables from the input data and reduce effects of noise or irrelevant variables[19]. The improved FP-growth algorithm is then used to discover the relations between the target attribute, i.e., non-performing loans and other attributes and rank the attributes. Finally, from the top factors discovered, the KPIs that influence the business goals were calculated using SSAS with multidimensional expressions (MDX) and their value, status and desired trends were obtained. To implement the FP-growth algorithm and test its performance, 2.6 GHz Intel Core i5, 8 GB 1600 MHz DDR3, OS X version 10.9.5 and Netbeans were used to compile the source code of the algorithms.

Data analysis: For the primary aims, Pearson's correlation coefficient applied which is used to measure the strength of a linear relationship between paired data[20]. Correlation coefficients matrix was calculated between target variable (Non-performing loan) and other variables are listed in Table 1. Table 1 shows strong correlation between Non-performing loan and frequency, age and salary.

RESULTS

The results reported in this study were obtained based on using two data sets which are a Kuwaiti Bank dataset covering a 6 years period from 2003-2008 and three datasets

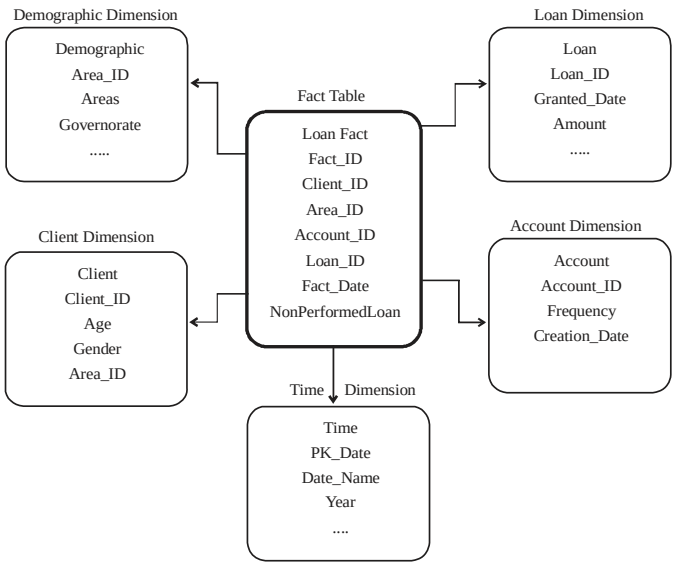

Fig. 2: Star schema for loans cube

Table 2: Attributes extracted from loan cube for feature selection

Table	Attribute	Description
Demographics	Salary	Salary of account owner
	Governorates	Governorate's name
	Area	Area name
Account	Frequency	Frequency of statements issuance
Loan	Duration	Duration of the loan
	Loan amount	Amount of money
	Non-performing loan	Status of loan-performing or non-performing
Client	Age	Client age
	Gender	Client gender

(Mushroom, Chess and Pumsb) were used for testing the performance of improved FP-growth.

Set business goals: The starting point to detecting the right KPIs is setting the right target. Business goals are highly individualized and strongly influenced by the size of the business, available resources, budget and many other factors that vary from business to business. The business goal for this study was to reduce non-performing loans.

Data preparation: The data was then made consistent, for example, by translating data in Arabic into English, replacing some attributes with their two-character codes and splitting the birthday field into two fields-gender and age.

Building data source: Part of any business intelligence project involves obtaining data from multiple sources to build a data warehouse and add cubes on top to get the best performance when slicing and dicing the data. In this case study, one data source was implemented as a database using server management studio.

Data pre-processing: At this stage, the age attribute was discretized into four categories (youth, adult, middle age and senior) and non-performing loans were divided into two categories (True and False). Similarly, other attributes like loan amount, loan duration and payments were also transformed into discretized attributes. Data aggregation involved grouping rows by year or month or by transaction type.

Building data warehouse: Data was extracted from the loan cube for feature selection based on the business goal, i.e., reducing non-performing loans. Loan cube implemented as a star schema to understand the data easily and increase performance and flexibility in navigating the data as shown in Fig. 2.

Feature selection: The attributes extracted for feature selection using the IG algorithm are listed in Table 2. The IDs and date attributes before feature selection have been excluded.

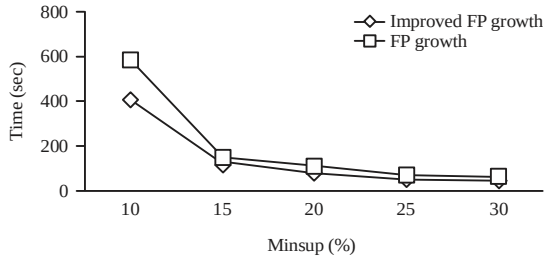

Fig. 3: Performance comparison experimental results with Mushroom

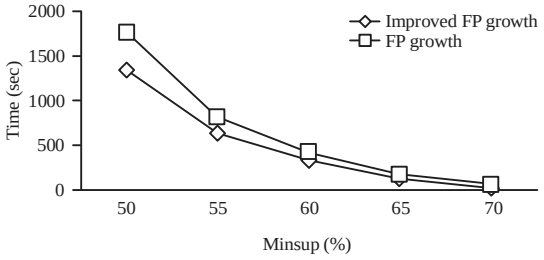

Fig. 4: Performance comparison experimental results with Chess

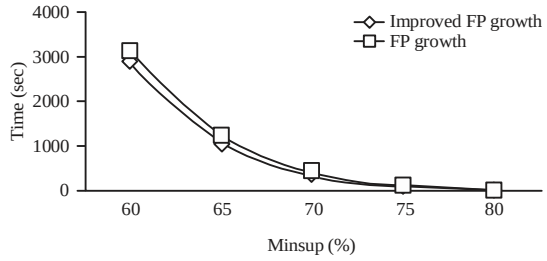

Fig. 5: Performance comparison experimental results with Pumsb

Table 3: Characteristics of three data sets chosen from the Frequent Item set Mining Dataset Repository

Datasets	Average transaction size	No. of transactions	Items
Mushroom	23	8124	119
Pumsb	74	49046	2113
Chess	37	3196	75

The IG algorithm eliminated four features and ranked the rest as follows: Gender, frequency, governorate's name, salary, duration and age.

Discovering association rules: For the experiment, three data sets were chosen from the frequent item set mining dataset repository[21]-Mushroom, Chess and Pumsb (Table 3).

The performance comparison results with different levels of minimum support are shown in Fig. 3-5. The results

```
1 [Young] -> [TRUE, Monthly]: 82%
2 [Low] -> [TRUE, Monthly]: 79.6%
3 [Young, Low] -> [TRUE, Monthly]: 79%
.
.
.
7 [Monthly] -> [Young, TRUE]: 73%
8 [Monthly] -> [TRUE, Low]: 70%
9 [Low] -> [Young, TRUE]: 67%
```

Fig. 6: Rules discovered by improved FP-growth algorithm

demonstrated that the improved FP-growth algorithm out-performed in terms of efficiency and scalability in time.

After demonstrating the out performance of the improved FP-growth algorithm, it was used to discover the association rules on loan cube (dataset from Table 2). The resulting rules (Fig. 6) determined the factors that influence the performance of non-performing loans.

Some samples of the rules discovered are illustrated in Fig. 6. Their interpretation is summarized as:

- IF "Age" = Young and "Frequency issuance" = Monthly then "non-performing loans" = True
- IF "Salary" = low and "Frequency issuance" = Monthly then "non-performing loans" = True
- IF "Age" = Young and "Salary" = low then "Frequency issuance" = Monthly and "non-performing loans" = True
- IF "Frequency issuance" = Monthly and Age = Young then "non-performing loans" = True
- IF "Frequency issuance" = Monthly and "Salary" = low then "non-performing loans" = True
- IF "Salary" = low and "Age" = Young then "non-performing loans" = True
- Calculating KPIs

The MDX to construct the non-performing loan KPI is shown in Table 4. The MDX expression delivers the basis for evaluating the progress towards the goal. In this MDX expression, the chosen graphic in SSAS will change according to the KPI status and trend.

DISCUSSION

The rules indicated that accounts with low salaries have more loan problems, i.e., a young client may be less relied upon to return loans. In addition, accounts with a monthly issuance of statements have more problems with loan repayments. The algorithm indicated no strong correlation between non-performing loans and attributes such as gender, loan duration and governorates. To evaluate and demonstrate

Table 4: MDX expression to construct non-performing loan KPI

Term	MDX expression
KPI name	[Non-Performing Loan KPI]
Value	[Measures]. [Non-Performing Loan]/[Measures]. [Fact Loan Count]
Goal	8
Status	Case when (KPI Value ("Non-Performing Loan KPI")) < 8 then-1 when (KPI Value ("Non-Performing Loan KPI")) > 8 then else 0
Trend	Case When Isempty(KPI Value ("Non-Performing Loan KPI")) Then Null When ([Time].[Year].prevmember, [Measures].[Non-Performing Loan]) < KPI Value ("Non-Performing Loan KPI ") Then -1 When ([Time].[Year].prevmember, [Measures].[Non-Performing Loan]) = KPI Value ("Non-Performing Loan KPI ") Then 0 Else 1 End

the feasibility of this approach, the proposed methodology was applied to another case study involving the use of credit cards. The results showed that the key factors affecting the performance of the credit card withdrawal KPI are related to the cube business goal of increased credit card usage.

Current KPI selection processes lack efficacy are complex and do not provide holistic approaches[22]. Moreover, KPIs that related to the business objective is hard to find[23]. This study proposed a method to detect KPIs based on evaluation attributes, using IG to determine the most important factors that affect KPIs. It then applied an improved FP-growth algorithm to discover the correlation between attributes. This method help to find the best rules with appropriate minimum support.

The main contribution of this proposed model was to enable organizations to detect and select the right KPIs, which are clearly linked with their business goals as well as discover the factors that influence the performance of these KPIs. Previous studies[24-27,7] have either focused on the application of a pre-defined list of KPIs extracted from a literature review and recommended by industrial professionals or from standard ISO. A method was proposed to discover KPIs but with the final decision being made by experts6, or discuss the importance and characteristics of KPIs but not how extracted them from data[28]. This study, however, presented a general model to detect KPIs without any prior experience in the domain.

CONCLUSION AND FUTURE RECOMMENDATIONS

The proposed framework can be used as a general method to detect KPIs. It allows decision makers to determine the most appropriate KPIs for their business goals, without the need of prior experience in the domain. The proposed framework is flexible enough to integrate new data sources that need to be plugged in for effective decision support and will thus be useful to decision makers. One of the limitations of this framework was that it was applied only on a limited number of nine attributes. Future studies should focus on using this approach with big data to detect KPIs handling more attributes and solving for issues such as detecting anomalies and predicting the deviation in KPIs.

SIGNIFICANCE STATEMENTS

- Detection the KPIs from historical data without prior experience in domain or predefined KPI list
- Using feature selection for ranking the factors that affect the performance of the KPIs
- Discover the correlation between attributes based on association rules

ACKNOWLEDGMENTS

The author are grateful to the supervisors: Dr. Ayman E. Khedr, Dr. Sherif Kholeif and Dr. Amira M. Idrees for useful discussions and comments for improving the ideas, results and theoretical framework.

REFERENCES

1. Al-Aqrabi, H., L. Liu, R. Hill and N. Antonopoulos, 2015. Cloud BI: Future of business intelligence in the cloud. J. Comput. Syst. Sci., 81: 85-96.
2. Spitzer, D.R., 2007. Transforming Performance Measurement: Rethinking the Way We Measure and Drive Organizational Success. AMACOM, New York, USA., ISBN: 9780814408919, Pages: 288.

3. Rusaneanu, A.E., 2014. Rules for selecting and using key performance indicators for the service industry. SEA-Pract. Applic. Sci., 2: 661-666.
4. Dumitrache, C., O. Kherbash and M. Mocan, 2016. Improving key performance indicators in Romanian large transport companies. Proceedings of the 13th International Symposium in Management: Management During and After the Economic Crisis, June 7, 2016, Timisoara, Romania, pp: 211-217.
5. Chaharsooghi, S.K., N. Beigzadeh and A. Sajedinejad, 2016. Analyzing key performance indicators of e-commerce using balanced scorecard. Manage. Sci. Lett., 6: 127-140.
6. Peng, W., T. Sun, P. Rose and T. Li, 2008. Computation and applications of industrial leading indicators to business process improvement. Int. J. Intell. Control Syst., 13: 196-207.
7. Ishak, A. and M. Sahak, 2010. Discovering the right key performance indicators in libraries. Proceedings of the 1st PERPUN International Conference and Workshop on Key Performance Indicators for Libraries, October 17-19, 2011, Malaysia, pp: 1-18.
8. Granberg, T.A. and A.O. Munoz, 2013. Developing key performance indicators for airports. Proceedings of the 3rd ENRI International Workshop on ATM/CNS, February 19, 2013, Tokyo, Japan.
9. Ross, R. and K. Ingo, 2014. Exploring customer specific KPI selection strategies for an adaptive time critical user interface. Proceedings of the International Conference on Intelligent User Interfaces, August, 2014, Israel, pp: 341-346.
10. Stefanovic, N., 2014. Proactive supply chain performance management with predictive analytics. Sci. World J., Vol. 2014. 10.1155/2014/528917.
11. Bouckaert, R., E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald and D. Scuse, 2013. WEKA manual for version 3-7-8. University of Waikato, New Zealand.
12. Sutha, K. and J.J. Tamilselvi, 2015. A review of feature selection algorithms for data mining techniques. Int. J. Comput. Sci. Eng., 7: 63-67.
13. Han, J., J. Pei and Y. Yin, 2000. Mining frequent patterns without candidate generation. Proceedings of the ACM SIGMOD International Conference on Management of Data, May 15-18, 2000, Dallas, TX., USA., pp: 1-12.
14. Pramod, S. and O.P. Vyas, 2010. Survey on frequent item set mining algorithms. Int. J. Comput. Applic., 1: 94-100.
15. Heaton, J., 2016. Comparing dataset characteristics that favor the Apriori, Eclat or FP-Growth frequent itemset mining algorithms. Proceedings of the Southeast Con, March 30-April 3, 2016, Norfolk, pp: 1-7.
16. Tanbeer, S., C. Ahmed, B.S. Jeong and Y.K. Lee, 2008. CP-Tree: A Tree Structure for Single-Pass Frequent Pattern Mining. In: Advances in Knowledge Discovery and Data Mining, Washio, T., E. Suzuki, K.M. Ting and A. Inokuchi (Eds.). Springer Science and Business Media, New York, ISBN: 9783540681243, pp: 1022-1027.
17. Pandya, M. and P. Trikha, 2013. A new tree structure to extract frequent pattern. Int. J. Emerging Technol. Adv. Eng., 3: 901-906.
18. Stacia, M., 2009. Business intelligence: Planning your first Microsoft BI solution. Microsoft TechNet Magazine. https://technet.microsoft.com/en-us/magazine/gg413261.aspx.
19. Chandrashekar, G. and F. Sahin, 2014. A survey on feature selection methods. Comput. Electr. Eng., 40: 16-28.
20. Zou, Q., J. Zeng, L. Cao and R. Ji, 2016. A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing, 173: 346-354.
21. Bayardo, R., 2004. Frequent itemset mining dataset repository. http://fimi.ua.ac.be/data/.
22. Stricker, N., A. Pfeiffer, E. Moser, B. Kadar and G. Lanza, 2016. Performance measurement in flow lines-key to performance improvement. CIRP Ann. Manuf. Technol., 65: 463-466.
23. Angoss Software, 2011. Key performance indicators, six sigma and data mining. White paper, Angoss Software Corporation, Canada, pp: 1-33.
24. Bai, C. and J. Sarkis, 2014. Determining and applying sustainable supplier key performance indicators. Supply Chain Manage.: Int. J., 19: 275-291.
25. Bimonte, S., E. Naoufal and L. Gineste, 2016. A system for the rapid design and implementation of personalized agricultural key performance indicators issued from sensor data. Comput. Electr. Agric., 130: 1-12.
26. Kucukaltan, B., Z. Irani and E. Aktas, 2016. A decision support model for identification and prioritization of key performance indicators in the logistics industry. Comput. Hum. Behav., 65: 346-358.
27. ISO., 2014. Automation systems and integration: Key Performance Indicators (KPIs) for manufacturing operations management-Part 2: Definitions and descriptions. Document No. ISO 22400-2.
28. Parmenter, D., 2010. Key Performance Indicators: Developing, Implementing and using Winning KPIs. 3rd Edn., John Wiley and Sons, New York.