Anomaly Detection in Role-based Workflows using Case-based Reasoning and the inverse Problem

Francis Ekpenyong[1], Stelios Kapetanakis[1], Miltos Petridis[2]

[1] School of Computing, Engineering and Mathematics, University of Brighton, Moulsecoomb Campus, Lewes road, Brighton BN2 4GJ, UK
[2] Department of Computing, University of Middlesex, The Burroughs, London NW4 4BT, UK
1{f.ekpenyong,s.kapetanakis}@brighton.ac.uk
2m.petridis@mdx.ac.uk

Abstract. This paper presents our work on how the Inverse problem (IP) can be applied to enhance Case-based Reasoning (CBR). It develops a novel model that extracts the capabilities of CBR in solving problems in domains where there is scarce or no mathematical foundations and combines it with inverse problem techniques (IPTs). IPTs have registered numerous successes in science and engineering fields since many important real-world problems can be solved via an application of IP. The areas of the problem space that are not represented in a case base will be created using Machine Learning (ML) algorithms whereby problems that are not captured by the current case base are generated and matched with a known case to ascertain the case genuineness. This is an inverse problem because it takes the output generated from the solutions of CBR probabilistically predicts relevant cases for the case base. It will be applied to a drilling technology workflow. We will present preliminary results which formulate parts of the forward problem. Sample data gotten from the relevant literature were simulated to test the model performance against some selected metrics. This approach will be useful not only for increasing the problem coverage of the case base but also in creating cases with rare solutions, improve the performance of an existing Case-based reasoning system by automatically generating synthetic cases hence boosting the CBR.

Keywords: Inverse Problem, Case-based Reasoning, Machine Learning, Business Process Workflows, Anomaly Detection

1 Introduction

The exponential growth of digitized technology has increased the dependency of human activities on its services, and in business, which in turn affects the mechanisms through which they create and capture value to earn profits and in some cases, result in various anomalies in business processes. Case-based reasoning (CBR) [1] as emerging methodology, have made significant contributions to the task of making predictions in business workflows [5], [9] and can be applied to business process as a support to

knowledge transfer [22] as most of the previously developed methodologies lack actual guidance on the process design, threatening the success of Business Process Relations (BPR) [2].

A score of basic countermeasures against anomalies in business processes have been proposed, sad though, these and other methods have proved to be insufficient or inap-propriate. Therefore, this research, will investigate the possibility of combining CBR which involves reuse of previous knowledge to solve new problem and Inverse Prob-lem(IP) technique, a task of employing process output information to recommend suit-able input settings for the process concerned, in detecting anomaly in human related workflows. We propose a generic method for deriving useful CBR system from busi-ness models, by developing a generic model which covers relevant information on the past business processes including the applied results. The model process is found by analysing all related design activity for each current activity.

A suitable knowledge representation and similarity measures will be developed and tested against a range of available ML algorithms and the most suitable one will be chosen, then the Genetic Algorithm (GA) will be applied on the derived model to create a more effective distribution of cases across the model which will predict sufficient case for the knowledge-base. The results of which will then be used as a case base for standard Case-based Reasoning process, and will be evaluated against a known episodic (real) data and human expert advice. The remainder of the paper is structured as fol-lows: section 2 discusses the motivation behind the research, report on the related liter-ature is reported in section 3, section 4 explains the methodology of the research work and section discusses the results and evaluation, section 6 talks on the conclusion and the future direction of the research.

1.1 Motivation

Drilling engineering is a process with increasing level of complexity due to various uncertainties surrounding the process which frequently lead to operational challenges. To deal with the complexity of the problems stemming from the vast amounts of pro-cess data generated and the considerable number of parameters involved, our approach extends a hybrid case-based reasoning and inverse approach with reasoning within a model of general domain knowledge. It is shown how the combined reasoning method empowers focused decision support for fault diagnosis and prediction of potential un-wanted events in this domain. The motivation behind this research is to advance a spe-cific computerized method for helping various drilling industry make a more useful and accurate informed decisions

2 Related Work

Various works have been carried out using Case-based reasoning (CBR) to support critical decision making in human related business workflows, either as a standalone or combination of techniques as well as inverse Problems as reported in [3], [4], [6], [13] Combination of Case based Reasoning and Process Mining to design and develop a

web based crisis management decision support system is reported in [14]. Some of the results of these researches show improved performance, but they still rely on existing cases and are not able to predict or generate sufficient cases for the Knowledge base.

A sizable number of different research efforts that employ CBR to improve the drilling operations [14] [16] point out that access to the data and information is a major problem in this the drilling domain.

Inverse theory on the other hand, involves a well-articulated set of mathematical techniques for reducing data to obtain useful information about the physical world. It is based on inferences drawn from observations of a physical phenomenon or environment to elicit features relating to the investigated phenomenon or environment [7]. It has registered numerous successes in science and engineering fields since many important real-world problems give rise to an Inverse problem (IP). These include medical imaging, non-destructive testing, oil and gas exploration, land-mine detection and pro-cess control. A couple of researches have also been carried out using the Inverse Problem. [4] [11] [19] [7] studied applications of methods from the theory of inverse problems to pattern recognition. However, most of these researches have been applied to solve science and engineering problems but little efforts have been made to combine Machine Learning techniques in this domain. A combination of techniques in data analytics can bring a promising result as evident by [8], This is also reported in [20] that explored the feasibility of using a case-based model as a tool to improve the usability of numerical models and to solve inverse, and constraint problems. Owing to the proposition of [8] and other researches, this research will combine CBR with Inverse Problem in predicting in drilling technology, which is a very complex system that is characterised by uncertain random information emanating from the drilling process.

2.1 Case Based Reasoning

Case-based reasoning is an Artificial Intelligence (AI) technique that supports the capability of reasoning and learning in advanced Decision Support System (DSS). It is a paradigm for combining problem solving and learning is analogous to problem solving that compares new cases with previous indexes cases. CBR provide two main functions: storage of new cases in the database through indexation module and searching the indexes cases with the similarities of new cases in case retrieval module [21].

The case-based reasoning methodology incorporates four main stages [1], [13], Retrieve: given a target problem, retrieve from the case memory, cases that are most relevant and promise to proffer solution to the target case.

Reuse: the solutions of the best; map the solution from the previous case to the target situation, test the new solution in the real world or perform a simulation, and if necessary.

Revise: the solution provided by the query case is evaluated and information about whether the solution has or has not provided a desired outcome is gathered.

Retain: After the solution has been successfully adapted to the target problem, the new problem-solving experience can be stored or not stored in memory, depending on the revise outcomes and the CBR policy regarding case retention.

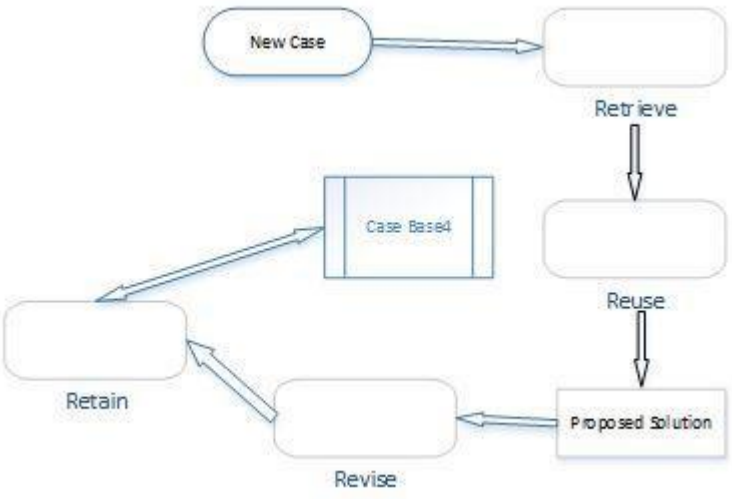

Fig. 1. CBR Cycle (Modified): Aamodt A, Plaza E, (1994)

Four knowledge containers within the CBR system as Vocabulary, Similarity measures, Adaptation knowledge and Cases [8]. [14] discussed CBR applied to problems in drilling engineering, showing that drilling engineering is an application domain in which the systematic storage and situation-triggered reuse of past concrete experiences provide significant support to drilling personnel at various levels. This informed the decision to apply the concept to this domain as well as presentation by [8] who emphasized strong innovations for integrating CBR with other reasoning modalities and computing techniques to more accurately model the knowledge available in a problem domain, compensate for lack of a productive or complete model of a problem do-main and compensate for the shortcoming of one approach by capitalizing on the strength of another.

2.2 Inverse Formulation

Inverse problems (also called model inversion) arise in many fields, where one tries to find a model that typically approximates observational data. Any inverse theory requirement is to relate physical parameter "x" that describes a model to acquire observations making up some set of data "y". Assuming there is a clear picture of the under-lying concept of the model, then an operator can be assigned a relation or mapping x to y through the equation F(x) = y; formulated in each vector space setting. The dependency between x and y (input and output parameters) can therefore be represented mathematically as follows:

$$Y_{k,\ldots}y_m = f_k(x_1 \ldots x_n) \qquad (1)$$

Fig. 2. Dependency between input and output parameters

Where i= 1 to n and n is the number of input parameters and yk is the output parameters; k ranges from 1 to m.

The problem of estimating x (inputs) from a measurement of y (observation) is a model of an inverse problem expressed in the same vector space setting. As such, the idea of estimating x from a measurement of y is a prototype of an inverse problem. Given by:

$$x_i = f(y_1, \ldots y_m) \qquad (2)$$

If the operator F is linear, the inverse problem is termed to be linear and the direct inverse is easy to find; otherwise it is a non-linear inverse problem, and termed ill-posed problem which poses considerable difficulties in solving. It appears that non-linear Inverse problems (also called model inversion) arise in many fields, where one tries to find a model that typically approximates observational data than the linear ones.

2.3 The Forward and Inverse Problems

The traditional forward function "f" (also called forward model) typically predicts a set of data that one is interested in. That is, f conducts simulations by entering different values for x and then examining the values of y that are generated as output. In a realistic situation, there may be many inputs (x) and outputs (y), and the nature of the function (f) may be complex.

Fig. 3. Forward Transformation

On the other hand, the inverse behaviour of the function f is to perform optimizations by setting a target value for y, then determining or estimating the values of x that result in the target value for y. Again, considering that the inputs (x) and outputs (y) may be large, which in effect can increase the complexity of the model resulting in elaborate search techniques and time consuming simulations and in worse cases becomes impractical in case of large model.

Fig. 4. Inverse Transformation

Contemporary techniques and achievements in the drilling process can also be visualised from two angles: analysis of drilling information (direct problems) and synthesis of drilling information (inverse problems).

Analysis of drilling information results in the set of features extracted from series or drilling reports of past successful or failed drilling operational cases while the synthesis of drilling information results in synthetic operational data designed by applying set of rules to the result of the analysis

2.4 Case Inversion

Inverse modelling refers to an attempt to derive a physical property of a model given a set of observations and in effect, the reverse operation to forward modelling, Considering the complexity of drilling process operation where frequent problems occur when drilling several kilometres through different geological formations. Each process may experience both similar and unfamiliar problems during the drilling operation, this could be much more challenging process than forward modelling because of the likelihood of deriving multiple solution. A way of addressing this non-uniqueness characteristic is to define some preferred characteristics for the solution. The inversion program can then find one or more solutions that not only reproduce the observations but satisfy our preferences. Figure 5 shows the proposed inversion workflow

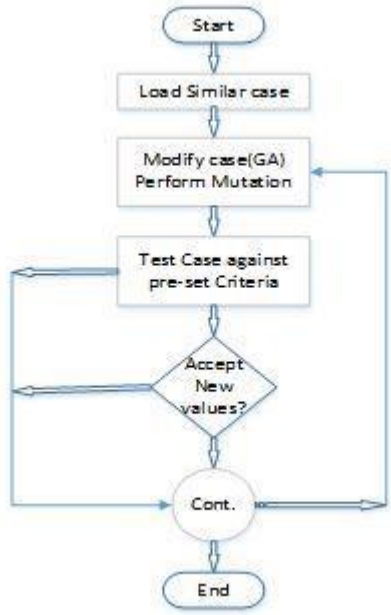

Fig. 5. Proposed inversion workflow

3 Methodology

In this research, a new drill process workflow is formulated, first as a forward problem, and then the Genetic Algorithm will be applied on the derived model to predict un-known (synthetic) cases to boost our knowledge base. This will be done through the following steps:

A knowledge mining model will be created, by developing a generic model which covers relevant information on the past drilling processes including the applied results. The model process is found by analysing all related design activity for each current activity. This will be evaluated by in two ways; by testing it against a simulated data and again by the service of human expert.

A suitable knowledge representation and similarity measures was developed and tested using the various algorithms and the best performed algorithm was selected. A new case will be compared, and similar case retrieved from the case repository, and if there is no match, then a new case will be formulated from the extracted features and the full case-based reasoning cycle performed.

After successfully determining the forward solution, the Genetic Algorithms will be applied on the derived model to create a more effective distribution of cases across the model which will predict sufficient case for the knowledge-base. The results of which will then be used as a case base for standard Case based-reasoning process which will be evaluated against a known episodic (real) data and human expert advice. The whole Inverse/CBR model concept is shown in figure 6.

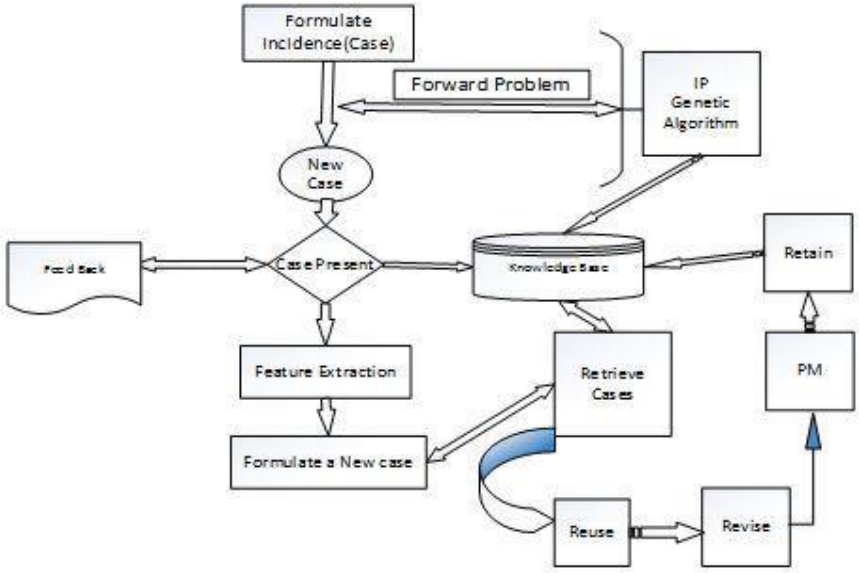

Fig. 6. CBR/Inverse Model

3.1 The Case Study

The above methodology will be centred on the drilling workflow and risk evaluation aspects of the operations process and will seek to enhance some valuable level of feed-back to the operations management system; the first level of feedback would be an effective and thorough drilling investigation following an occurrence of incident, which would lead to an evaluation and improvement of the failed system that caused the incident. The next feedback level would be dedicated on ensuring that valuable knowledge in the form of preventive strategies and incident investigation reports are made available and useable for new project planning processes. The second level feedback implementation would smoothen the transfer of knowledge across projects and learning from past drilling project failures.

The major compositions of the CBR approach are:

- a detailed drilling knowledge representation scheme that can be used to capture and abstract key drill operational knowledge from previous drill operations.
- a case retrieval mechanism based on customized similarity scoring functions.
- case identification adaptations that facilitate modifications of retrieved cases and assimilation of all significant cases.

3.2 Experimental dataset and Data pre-processing

Real time data can be gotten in drilling operations via sensors transmitted and translated to a suitable format that can be easily interpreted by the drilling crew. In addition to

real-time data, documents such as daily drilling reports and end of drilling reports are produced for every single well in which solutions for most of the past problems are expressed. Descriptions of situations in these reports contain the occurring problems and their proposed solutions. The later forms the basis our acquisition; datasets were gotten from reports where some relevant features were selected to form our sample dataset.

We performed some experiments to determine the goodness of the created model and used some statistical methods to make estimate on the accuracy of the models that we created. To test the performance of the model, 10-fold cross validation was applied to the dataset and all the four classifiers were experimented. The datasets were partitioned into training set for training the model and test set for testing the model. For the choice of samples in training the percentage of each class was balanced statistically, while for testing portions, the percentage of each class in each portion is preserved. The training and the testing portions were further adjusted in the ratio of 10:90, 30:70, 50:50, 70: 30 and 90:10.

To determine which algorithm will be ideal for our model, some selected ML algorithms were used. Two simple linear algorithms; Linear discriminant analysis (LDA) and logistic regression (LR), and two nonlinear k-Nearest Neighbours (kNN) and Sup-port vector machines (SVM) algorithms were selected. LDA and LR are widely used multivariate statistical methods for analysis of data with categorical outcome variables. Both are appropriate for the development of linear classification models associated with linear boundaries between the groups. While the kNN, determines the nearest k training instances to a target instance. SVM on the other hand, attempts to find a hyper-plane separating the different classes of the training instances, with the maximum error mar-gin.

We reset the random number seed before each run to ensure that the evaluation of each algorithm is performed using the same data splits. This is to ensure that the results are directly comparable. We then later run the different model directly on the validation set and summarize the results as a final accuracy score, a confusion matrix and present the outcome of the classification.

4 Results and Evaluation

This section presents the experimental results on how the selected algorithms perform on the experimental drill dataset. The algorithms were Linear discriminant analysis (LDA), logistic regression (LR), kNN, and SVM. The choice of these was based on the fact that they are the most popularly used algorithms. The results show their performances derived from 10-fold cross validation of the training and test data. The results are shown from figure 7, and figure 8.

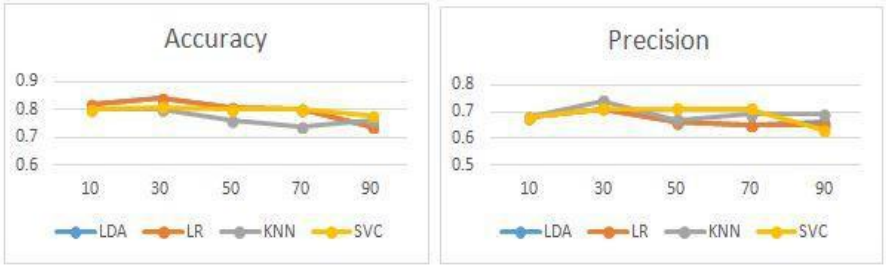

Fig. 7. Accuracy and Precision for the Classifiers

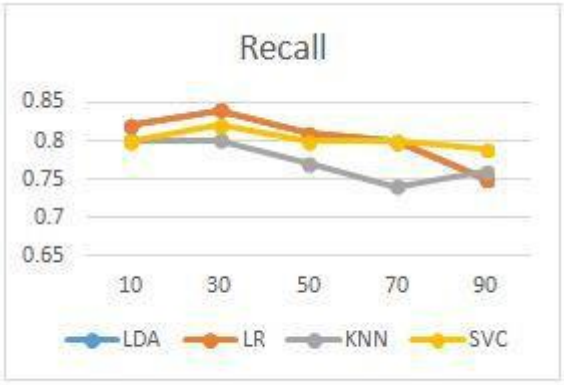

Fig. 8. Recall for the Classifiers

From the results, the variation of the percentage ratio for the training dataset has significant effect on the behavioural pattern of the various performance metrics used. The result also shows that there is no significant difference between the two linear algorithms (LR and LDA), while the algorithms have their overall best performances at 30:70 data validation/train ratio. The performances also seem to decrease with increase

in the validation data. kNN performs better in terms of precision. It can also be seen that LDA and LR show overall better performances.

5 Conclusions and future work

CBR is a recent methodology compared to many other computer science branches, especially in the drilling industry. This paper proposed a hybrid model of case-based reasoning and Inverse problem approach to identifying anomalies in business pro-cesses. Our experimental results formulate a preliminary study in "creating" the for-ward problem. The inverse problem was used to generate synthetic cases for the case base based on genetic algorithms. The GA initial population in our work was used as cases retrieved via CBR. The evaluation of four selected ML algorithms were done and their accuracy reported which will guide us on the ideal choice in building our model.

However, by capturing such experiences in a new model that combines Case-Based Reasoning technique and Inverse Problem, companies can learn lessons about actual challenges to management assumptions, adequate project preparedness and planned execution and be able to leverage that knowledge for efficient and effective management of future similar operations.

It is also pertinent to point that the dataset used was characterised with linearity, incomplete information, fuzziness and uncertainty owing to the legal, corporate and societal impact of having confidential information such as this drill operations data in public domain. This impact negatively on exploring the experiment in greater depth. To prove the efficiency of our method, we used synthetic simulated data in evaluating their performances.

For future work, the plan will be to model some of the uncertainties to create a knowledge pool of distinct types of drilling patterns and apply CBR in computing the similarities and characteristics of the case using controlled experiment and the result tested against real drilling operations data.

References

1. Aamodt, A., Plaza, E. (1994), Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications. IOS Press, Vol. 7: 1, pp. 39-59.
2. Selma, L., Farhi, M. (2003), Case Based-Reasoning as a Technique for Knowledge manage-ment in Business Process Redesign
3. Abutair, H. Y. A., Abdelfettah, B. (2017) Using Case-Based Reasoning for Phishing Detec-tion. Procedia Computer Science 109, pp. 281-288.
4. Alvarez Acevedo, N. I., Roberti, N. C., Silva Neto, A. J. (2004), A one-dimensional inverse radiative transfer problem with time-varying boundary conditions. Inverse Problems in Sci-ence and Engineering 12 (2) (2004), pp. 123-140.
5. Kapetanakis, S., Petridis, M.: Evaluating a Case-Based Reasoning Architecture for the In-telligent Monitoring of Business Workflows, in Successful Case-based Reasoning Applica-tions-2, S. Montani and L.C. Jain, Editors. 2014, Springer Berlin Heidelberg. pp. 43-54.

6. Rogge-Solti, A., Kasnec, G. (2013) Temporal Anomaly Detection in Business Processes Business Process Management Volume 8659 of the series Lecture Notes in Computer Science pp. 234-249
7. Sever, A. (2015), An inverse problem approach to pattern recognition in industry. Applied Computing and Informatics 11.1 1-12.
8. Marlin, C., Sqalli, M., Rissland, E., Munoz-Avila, H., Aha, D. (2002), Case-Based Reason-ing Integrations. AI Magazine, no. 1, pp. 69-86, 2002
9. Kapetanakis, S., Petridis, M., Knight, B., Ma, J.,Bacon, L.: Providing Explanations for the Intelligent Monitoring of Business Workflows Using Case-Based Reasoning, in workshop proceedings of ExACT-10 at ECAI 2010, Lisbon, Portugal (2010)
10. Filippoupolitis, A., Loukas, G., Kapetanakis, S. (2014). Towards real-time profiling of human attackers and bot detection. In Proceedings of CFET 2014: Cybercrime Forensics Edu-cation & Training, Canterbury, UK.
11. Klaus, M., Tarantula. A. (2002), 16 Probabilistic approach to inverse problems. International Geophysics 81 (2002): 237-265.
12. Richter, M. M., Aamodt, A. (2006), Case-based reasoning foundations. The Knowledge En-gineering Review, Vol. 20:3, 203–207
13. Ralph Bergman (2016), Towards Case-Based Adaptation of Workflows. 18th International Conference, ICCBR 2010, Alessandria, Italy, July 19-22, 2010 Proceedings.
14. Shokouhi, S. V., Aamodt, A., Skalle, P. (2010) Applications of CBR in Oil Well Drilling: A General Overview. 6th IFIP TC 12 International Conference on Intelligent Information Processing (IIP), Oct 2010, Manchester, United Kingdom. Springer, IFIP Advances in Information and Communication Technology, AICT-340, pp.102-111, 2010, Intel. Information Processing V.
15. Triki, ., Ben Saus, N. B., Dugdale, J., Hanachi, C. (2013) Coupling Case-based reasoning and process mining for a web based crisis management decision support system. 2013 work-shops on enabling Technologies: Infrastructures for collaborative Enterprise.
16. Skalle P., Aamodt A. (2005) Knowledge-Based Decision Support in Oil Well Drilling. In: Shi Z., He Q. (eds) Intelligent Information Processing II. IIP 2004. IFIP International Federation for Information Processing, vol 163. Springer, Boston, MA
17. Kapetanakis, S., Filippoupolitis, A, Loukas, G., Al Murayziq, T. S. (2014). Profiling cyber attackers using Case-based Reasoning. In proceeding of the 19th UK Workshop on Case-Based Reasoning (UKCBR 2014), 9th December 2014, Cambridge, UK, pp.39-48
18. Roth-Berghofer, T., Recio-Garcia, J. A., Sauer, C. S., Bach, K., Klaus-Dieter, A., Diaz Agudo, B., Gonzalez Calero, P. A. (2012), Building Case-based Reasoning Applications with myCBR and COLIBRI Studio. Proceedings of UKCBR workshop on case based Reasoning pp. 71-82
19. Ulrych, T. J., Sacchi, M. D., Woodbury, A. (2001), A Bayes tour of inversion: a tutorial. Presented a tutorial on the Bayesian approach to the solution of the ubiquitous problems with some special emphasis on the concepts and approach to solving inverse problems.
20. Woon, F. L., Knight, B., Petridis, M., Patel, M. (2005), CBE-conveyor: a case-based reasoning system to assist engineers in designing conveyor systems. International Conference on Case-Based Reasoning. Springer Berlin Heidelberg, 2005
21. Kang, Y. B., Krishnaswamy, S., Zaslavsky, A. (2014), A Retrieval Strategy for Case-Based Reasoning Using Similarity and Association Knowledge. IEEE Transactions on Cybernet-ics, Vol. 44, (4), April 2014
22. Kapetanakis, S., Petridis, M.,Knight, B.,Ma, J.,Bacon,L. : A Case Based Reasoning Approach for the Monitoring of Business Workflows, 18th International Conference on Case-Based Reasoning, ICCBR 2010, Alessandria, Italy, LNAI (2010)