World Applied Sciences Journal 32 (3): 451-458, 2014
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2014.32.03.230

Case Retrieval Phase of Case-Based Reasoning Technique for Medical Diagnosis

[1]*Syed Saood Zia,* [2]*Pervez Akhtar,* [3]*Tariq Javid Ali Mughal and* [4]*Idris Mala*

[1]Faculty of Engineering, Sciences and Technology (FEST),
Hamdard University, Karachi, Sindh, 75600, Pakistan
[2]Electrical and Power Engineering, Pakistan Navy Engineering College,
National University of Science and Technology, Karachi, Sindh, 75350, Pakistan
[3]Department of Electrical Engineering, HITEC University, Taxila Cantt, Punjab, 47070, Pakistan
[4]Faculty of Engineering, Sciences and Technology (FEST),
Hamdard University, Karachi, Sindh, 75600, Pakistan

Abstract: In the field of medical sciences the technique of case based reasoning (CBR) offers a basis of computer aided decision support applications which ensure the accuracy of decisions during diagnosis and treatment phases of patient care. Physicians employ CBR technique to solve the new case by retrieving the preceding similar cases that are stored into the case repository. Aim of this paper is to work on case retrieval phase of the CBR technique which is applied on Breast Cancer data sets from UCI Machine Learning Repository. Results will be generated using myCBR tool.

Key words: CBR · Case Retrieval · Case Repository · Breast Cancer · myCBR

INTRODUCTION

Estimated calculations predict that cancer will affect 50% more people by the year 2020. Rate of survival from the disease is 89% in US and 76% in Europe. In developed countries this rate is progressively low [1]. Domain knowledge and health information play a vital role in making the cancer control campaign more effective. This would aid clinicians for making quality decisions [2].

Accuracy of decisions in diagnosis and treatment of diseases require practical experience of years. In this regard medical domain knowledge is an asset for medical practitioners which enable them to understand the complexity of the diseases [3]. Medical domain knowledge is the experiences gained from the past situations that include suggesting treatment and prescribing medication etc. [4]. In the beginning of medical practices, medical domain knowledge helps clinicians in dealing a particular situation. Cognition based model of case based reasoning (CBR) technique is very suitable for medical diagnosis [5].

The technique of CBR which comes under the Artificial Intelligence area deals with the knowledge that is derived from the similar cases [6]. In healthcare CBR technique present an ideal way out to handle complex situations during the diagnosis and treatment phase [7]. Medical practitioners utilize CBR technique to solve new cases by retrieving the preceding similar cases that are stored in to the case repository [8]. An important phase of the CBR application is to retrieves those cases from the case repository that are closely related to the new case.

In this paper we focus our attention towards explaining the case retrieval phase of the CBR technique which is applied on Wisconsin Breast Cancer data sets from UCI Machine Learning Repository. Results will be generated using myCBR tool.

This paper is organized as follows: Section 2 outlines the literature review. Section 3 shows the CBR cycle. Section 4 focuses on the Case retrieval phase of CBR technique. Section 5 presents on the different case retrieval algorithms that are used to retrieve the most

Corresponding Author: Syed Saood Zia, Faculty of Engineering, Sciences and Technology (FEST),
Hamdard University, Karachi, Sindh, 75600, Pakistan.
Cell: +92-346 276 7788.

similar cases from the case repository. Section 6 provides an overview of the CBR tool myCBR. Section 7 gives brief information of the UCI breast cancer dataset. Section 8 shows the implementation of Breast Cancer dataset using myCBR tool. Finally, Section 9 concludes this study and identifies some future directions of this research.

Literature Review: Categorizing medical data sets for the purpose of analyzing the specific situation is a complex task. Current survey shows that there is an enhanced progress on developing multiple phases of the CBR system.

The beginning of work on CBR systems were initiated in Yale University by Roger Schank and his colleagues at university of Yale in the year of 1977 [9]. They have focused their attention on making a system that comprises of dynamic memory for the purpose of solving problems by considering past experiences and their results [10]. In medical domain CBR applications are famous due to its cognitive based approach of solving problems. Currently CBR technique is employed in designing the applications that would aid medical practitioners during practices [11]. The foremost factor that ensures performance of CBR systems is a proficient way to retrieve cases from the case repository [12].

Similarity measurement approach is being used to extract closely related cases from the case repository. For executing similarity measurement, local and global similarity measurement is applied which is based on feature values of the cases. Nearest neighbor algorithm is generally used in most CBR systems for the purpose of retrieving closely related cases. The process of nearest neighbor algorithm for retrieving the cases is divided into two steps. Initially the related cases are selected by listing the cases into the case repository. After that, similarity measurement approach is applied for new case [13].

In CBR [14] utilized the comparison of different similarity calculations methods. These similarity measurement procedures were classified into three categories that are Case-biased similarity method, Query-biased and Equally-biased similarity calculation methods. Performances of these methods are examined by using different sample sizes.

Nearest neighbor algorithm is being used by [15] to retrieve closely related cases and accredited that the procedure is not proficient when the case base are too large.

The case base similarity calculation framework was proposed by [16]. This system was proficient enough to perform similarity calculation between different cases. Drawback of the system is that it is less capable in the adaption phase of the problem solving.

Grey incidence theory based framework was proposed by [17] had found this system efficient in designing banking sector applications.

The research work shown above for the similarity calculation is executed by means of nearest neighbor algorithm and it was found that in case of large case base, this method is inefficient.

This research paper presents the case retrieval phase of the CBR technique using myCBR tool which retrieves cases in more efficient manner [18]. The average weighted Euclidian distance method is used to compute the distance between the attributes of the input case with the stored cases. Attributes of the cases are assigned as weights on the basis of their importance by the domain medical experts. The distance measure range of the cases can be normalized into 0 to 1. After calculating the distance, compute the similarity as, when we have the distance between the input and stored cases is 0 the similarity is 100%.

CBR Cycle: In the medical field, case based reasoning approach is experiencing quick development due to its cognitive based model of solving problems. CBR technique uses past experiences to solve new cases with similar attributes.

The fundamental feature of the CBR systems is an inter-thread procedure of four phases which can be explained by a schematic cycle namely retrieval, reuse, revise and retain.

Retrieval phase is an initial step which inquires about previous experiences that are similar to the new case. In this phase most similar cases will be retrieved from the case repository.

Reuse phase is the second step which is responsible in suggesting a solution for the new case from the available solutions of the cases that were retrieved from the case repository. Solution proposed into the reuse phase will be then revised by an expert (either human or machine).

Once the solution is revised by the experts it has to be determined whether to keep this new solution in the case repository in order to facilitate the future diagnosis of new case. Keeping the proposed solution in case repository as a new case is the retain phase of this cycle [12, 19-21].

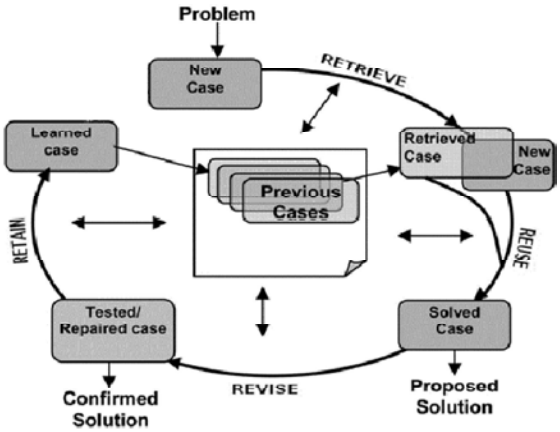

Fig. 1: CBR cycle [19].

Case Retrieval Phase of CBR: Retrieval phase is deemed to be fundamental part of the CBR technique. In retrieving phase an efficient way to extract the most similar cases is being used. The case retrieval process identifies to extract out the most suitable cases from the case repository which is closely related to the new given case.

Retrieval phase is further divided into four sub tasks [22].

Features Recognition: In order to accomplish an efficient retrieving process there is a need of setting of some standards or criterion for the selection of cases. This would determine which case is considered to be most suitable for the retrieval purpose. The domain expert or clinicians who retrieves the case is also influenced this selection criterion because it is also important that what a retriever is searching for in the case repository either the retriever is searching for few similar attributes of the new case and stored cases or they look for the entire similar case with the new case.

Case Inspection: Retrieval phase of the CBR system depends in the memory representation and indexing method of the stored cases into the case repository.

Similarity Measurement: Cases that are to be retrieved from the case repository were initially matched with the new case on the basis of the similarity in feature values.

Retrieve Similar Cases: Once the similarity computation has done, the most similar cases retrieved from the case repository that is closely matched with the new given case.

Case Retrieval Algorithms: Multiple algorithms and techniques are employed in CBR systems for the purpose retrieving cases from the case repository. Basis of these algorithms is similarity metric that allows comparison between current problem and previously stored cases of the case base. K-nearest neighbor, decision tree, Euclidian distance and their derivatives are the few techniques that were used more repeatedly for similarity computation purpose [21].

Nearest-Neighbor Retrieval: Nearest- neighbor algorithm is used to retrieve cases, based on weighted summation of the attributes of the cases. In case of more than one retrieved cases that case will be retrieved from the case base whose weighted summation of attributes is greater than other cases. Importance of the attributes depends on weights that are assigned by the domain experts [21].

Inductive Approaches: Inductive methodology for case retrieval is applied to find out the structure of case base and decides which attributes are relatively important and discriminated amongst related cases. The consequential case base structure will provide case retriever a compact space to search [21].

Knowledge-Guided Approaches: The technique of knowledge-guided for retrieving cases utilizes particular domain knowledge to find out those attributes of a case that could be relatively important in future retrieval of that case. Relative importance of attributes varies with the variation in situations [21].

Euclidian Distance: In order to compute distance between the attributes of two cases [23]. Attributes of the cases are assigned as weights on the basis of their importance by the domain experts. The range of the distance measure of the two cases can be normalized into 0 to 1. The average weighted method also incorporates with the Euclidian distance for assigning the priority of the feature values or attributes of the case. So the Weighted Euclidian distance formula for calculating the distance between the input case and the pile up cases that stored in the case library is shown in eq. 1.

$$d(\mathrm{C_N},\mathrm{C_O}) = \frac{\sum_{i=1}^{n} \mathrm{W_i} \times \sqrt{\left|\frac{C_{Ni} - C_{Oi}}{C_{Max\,i}}\right|^2}}{\sum_{i=1}^{n} \mathrm{W_i}} \quad (1)$$

Where

C_N = New referred case from clinicians
C_O = Previously stored case in a case repository
C_{Max} = Maximum value selected from the new referred case or previously stored case for converting it in normalized form.
N = Attributes in each case.
I = Is an individual or signal attribute.
W = Weight of each attribute. These weights determine the importance of each attributes and are assigned by field experts.

Once the distance between the new input case and the stored cases are computed then apply similarity measurement function that will shows the most similar cases with the new input case. The calculation performs for* computing similarity measurement function shows in eq. 2 and eq. 3.

$$\text{Sim}(C_N, C_O) = [1-d(C_N, C_O)] * 100 \quad (2)$$

$$\text{Sim}(C_N, C_O) = \left[1 - \frac{\sum_{i=1}^{n} W_i \times \sqrt{\left|\frac{C_{Ni} - C_{Oi}}{C_{Max\,i}}\right|^2}}{\sum_{i=1}^{n} W_i}\right] * 100 \quad (3)$$

In this paper, we use Euclidian distance method for retrieving the most similar cases from the case repository using myCBR tool. myCBR is a CBR application tool whose center of attention is the retrieval phase of CBR cycle which based on similarity metrics.

Overview of myCBR Tool: myCBR is an open-source plug-in tool that use the open-source editor Protégé [4]. The environment of the open-source editor Protégé is plug-and-play that supports rapid prototyping and application development [10]. Protégé based on object-oriented approach for defining classes and attributes. It also manages the instances of these classes and myCBR interprets these instances as cases [23].

myCBR provides a rapid prototyping approach for constructing CBR application based on similarity metric function used for retrieval of cases from the case library. myCBR provides user friendly GUI environment used for modeling various kinds of attribute specification similarity measures and evaluate the retrieved results [24].

Table 1: Wisconsin Breast Cancer Dataset

Attribute Name	Possible Values
Case Description Attributes	
CaseID	Id #
Clump Thickness	1 – 10
Uniformity of Cell Size	1 – 10
Uniformity of Cell Shape	1 – 10
Marginal Adhesion	1 – 10
Single Epithelial Cell Size	1 – 10
Bare Nuclei	1 – 10
Bland Chromatin	1 – 10
Normal Nuclei	1 – 10
Case Solution Attribute	
Class [2 Benign,4 Malignant]	[2, 4]

Breast Cancer Dataset: Breast cancer can cause the death rate to increase amongst woman. Breast cancer is classified into non-cancerous tissue (benign) or cancerous tissue (malignant). This study identifies the diagnosis of breast cancer tumor that either it is non-cancerous (benign) or cancerous (malignant) [26]. Case based reasoning technique is used as an inference mechanism. The datasets of breast cancer used for constructing the case library obtained from the University Wisconsin Hospitals, Madison from Dr. William H. Wolberg [25]. Total number of cases in the breast cancer dataset is 699 in which 34.5% cases are relate to cancerous class and 65.5% cases related to non-cancerous class. There are 54 cases which have missing values so these cases have been discarded from the case library [26]. Table 1 shows attributes and their possible values [25].

RESULTS

Using myCBR tool, the following steps is taken by the medical practitioners to examine the breast cancer patients.

Generation of Case Representation: Stored Cases in the case repository can characterize different types of knowledge and represent that knowledge in different representational format like ontology, frames etc. The representation of cases can be comprised into two set of attribute-value pairs, first one gives case description attributes information while other shows the solution features. myCBR provides the easiest way to represent the cases to the medical practitioners. The medical practitioners import the case library in the form of *.csv file into the myCBR tool. Figure 2 shows to import the *.csv file in myCBR tool.

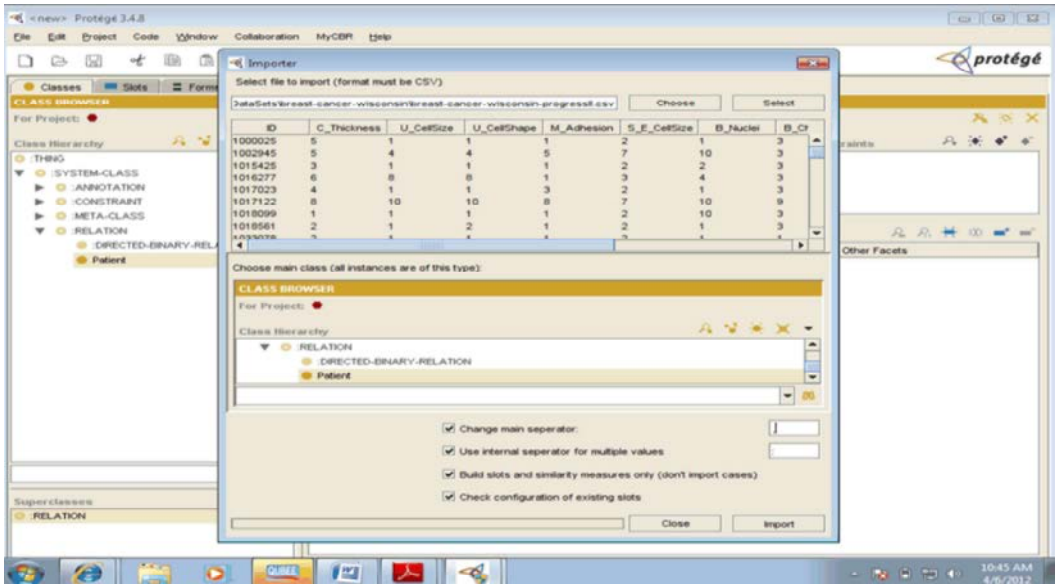

Fig. 2: Import CSV file in myCBR tool

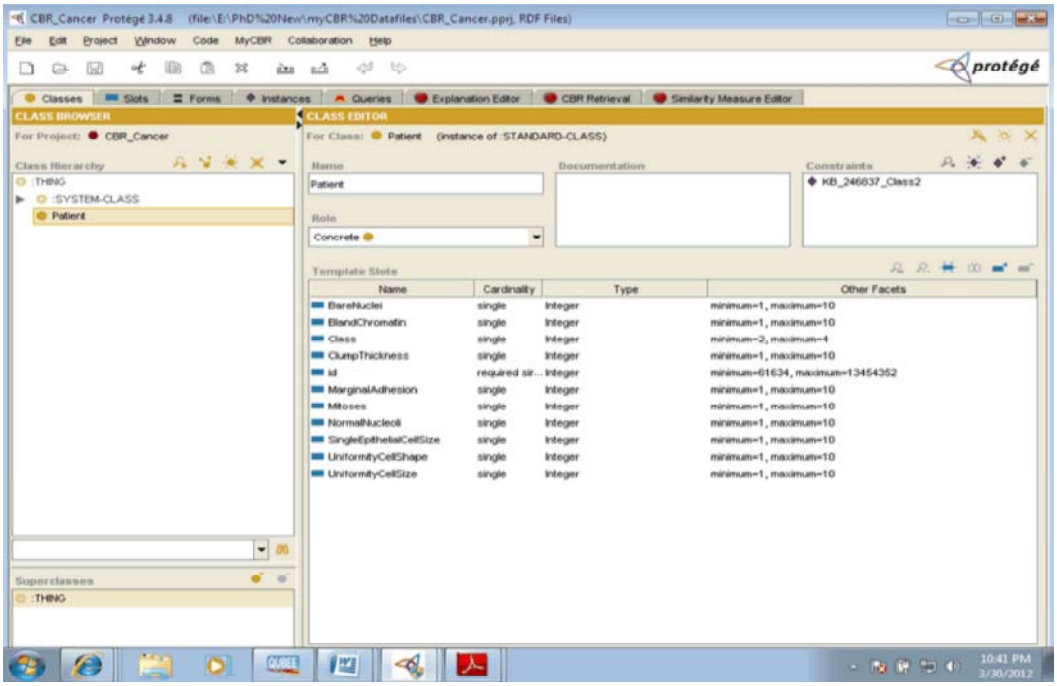

Fig. 3: Case representation in myCBR

After importing the case library in myCBR, it will generate the case representation [Figure 3] that shows the case description and case solution attributes value.

Modeling of Similarity Measure: In similarity measure editor, the medical practitioners assigned the weights to the attributes of disease that starts from [1 to 10] i.e. 1 is minimum and 10 is the maximum weight of the feature value. Figure 4 shows the similarity measurement editor in myCBR for assigning the weights of each feature values.

Case Retrieval Functionality: In Case retrieval process of myCBR input the new pathological report values of the patient i.e. the parameters of new case defined; the more similar cases are retrieved on the basis of average

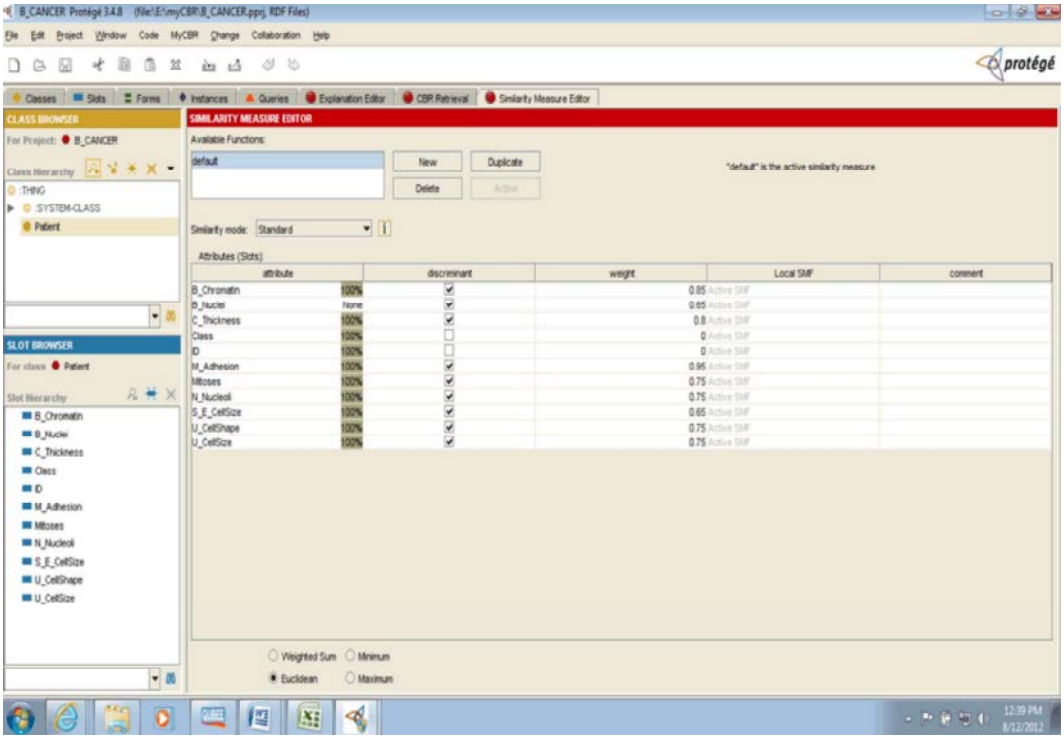

Fig. 4: Assigning weights for attributes

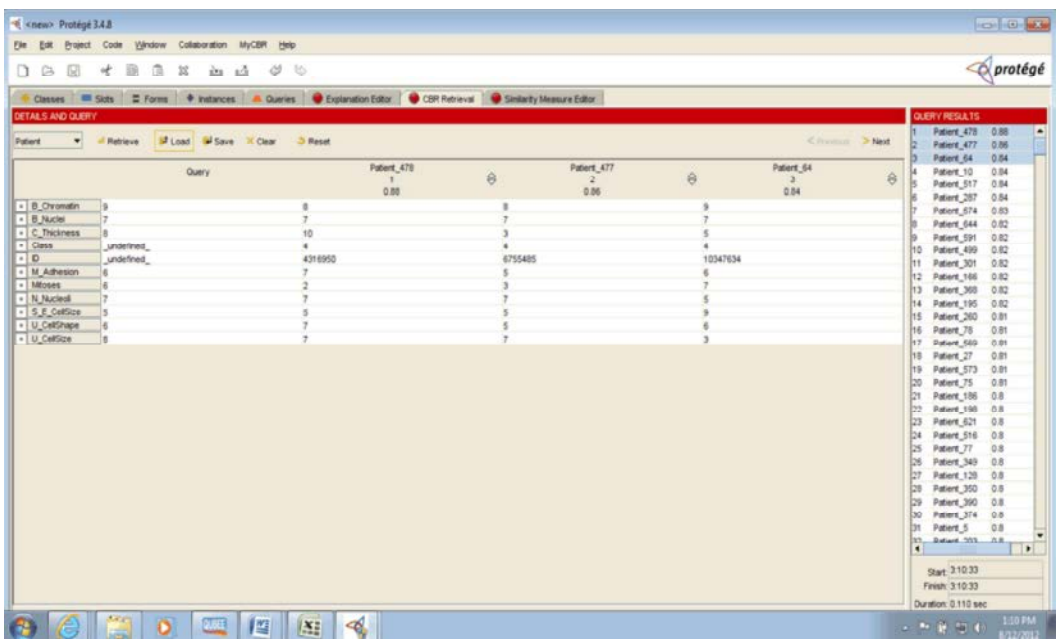

Fig. 5: Retrieve Similar Cases

weighted Euclidian distance method applied. Figure 5 shows to retrieve the most similar cases from the case library that is closely related to the attributes of input case values.

CONCLUSION AND FUTURE WORK

Case based reasoning is considered to be an ideal cognition based model used to develop reasoning

analogy. In medical domain CBR technique is employed to facilitate clinicians during diagnosis and treatment phase of patient care. This paper presents a brief overview of case based reasoning (CBR) technique and its schematic cycle. For the purpose of similarity computation different case retrieval algorithms have been discussed in this paper.

This study introduces an average weighted Euclidian distance method for computing the similarity between cases. Perform data analysis on the Wisconsin Breast Cancer data sets from UCI Machine Learning Repository and for generating results myCBR tool is being used. myCBR tool focuses on case retrieval phase of case based reasoning technique that used for retrieval of cases from the case library.

The web based CBR application would be a forthcoming piece of work that would be helpful for the young physician for decision making. Database management will be used to store complex medical case in form of table with relation to other tables. It also represents in graphical form that contain decision and also provides recommendation for the specific problems.

REFERENCES

1. Siegel, R., D. Naishadham and A. Jemal, 2012. Cancer statistics 2012. CA: a cancer Journal for Clinicians, 62(1): 10-29.
2. Sinha, R.K., M.M. Pai, M.S. Vidyasagar and B.M. Vadhiraja, 2010. A novel knowledge base decision support system model for breast cancer treatment. Sri Lanka Journal of Bio-Medical Informatics, 1(2): 97-103.
3. Lin, R.H. and C.L. Chuang, 2010. A hybrid diagnosis model for determining the types of the liver disease. Computers in Biology and Medicine, 40(7): 665-670.
4. Swe, T.M.M. and N.S.M. Kham, 2010. Case-based medical diagnostic knowledge structure using ontology. In Computer and Automation Engineering (ICCAE), 2010. The 2nd International Conference on 1: 729-733. IEEE
5. Gierl, L., M. Bull and R. Schmidt, 1998. CBR in Medicine. In *Case-Based Reasoning Technology* (pp: 273-297). Springer Berlin Heidelberg.
6. Salem, A.B.M., 2007. Case based reasoning technology for medical diagnosis. World academy of science, Engineering and Technology, 31: 9-13.
7. Bergmann, R., W. Wilke, K.D. Althoff, S. Breen and R. Johnston, 1997. Ingredients for developing a case-based reasoning methodology. In Proceedings of the 5th German Workshop in Case-Based Reasoning (GWCBR'97), LSA-97-01E, University of Kaiserslautern. pp: 49-58.
8. Zhuang, Z.Y., L. Churilov, F. Burstein and K. Sikaris, 2009. Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. European Journal of Operational Research, 195(3): 662-675.
9. Schank, R.C., 1983. Dynamic memory: A theory of reminding and learning in computers and people. Cambridge University Press.
10. Castro, J.L., M. Navarro, J.M. Sánchez and J.M. Zurita, 2011. Introducing attribute risk for retrieval in case-based reasoning. Knowledge-Based Systems, 24(2): 257-268.
11. Bareiss Jr, E.R., 1988. Protos: aunified approach to concept representation, classification and learning.
12. Lopez De Mantaras, R., D. McSherry, D. Bridge, D. Leake, B. Smyth, S. Craw and I. Watson, 2005. Retrieval, reuse, revision and retention in case-based reasoning. The Knowledge Engineering Review, 20(03): 215-240.
13. Lodhi, I.Y., K. Hasan, U. Hasan, N. Mahmood, T. Yoshida and M.A. Anwar, 2003. Optimizing retrieval process and using neural networks for adaptation process in Case Based Reasoning Systems. In Multi Topic Conference, 2003. INMIC 2003. 7th International, pp: 354-360. IEEE.
14. Gu, M., X. Tong and A. Aamodt, 2005. Comparing similarity calculation methods in conversational CBR. In Information Reuse and Integration, Conf, 2005. IRI-2005 IEEE International Conference on. pp: 427-432. IEEE.
15. Zhi-Ying Zhang, Jian-Wei Wang, Xiao-Peng Wei and Wen-Jing Yu, 2008. A model for case retrieval based on ANN and nearest neighbor algorithm. Machine Learning and Cybernetics, 2008 International Conference on 1: 142 - 147. IEEE.
16. Osborne, H.R. and D.G. Bridge, 1996. A case base similarity framework. In Advances in Case-Based Reasoning (pp: 309-323). Springer Berlin Heidelberg.

17. Mi, C., H. Qian, S. Liu and Z. Chang, 2008. Study on case retrieving in case-based reasoning based on grey incidence theory and its application in bank regulation. In Fuzzy Systems, 2008. FUZZ-IEEE 2008(IEEE World Congress on Computational Intelligence). IEEE International Conference on (pp: 1530-1533). IEEE.
18. Stahl, A. and T.R. Roth-Berghofer, 2008. Rapid prototyping of CBR applications with the open source tool myCBR. In Advances in Case-Based Reasoning (pp: 615-629). Springer Berlin Heidelberg.
19. Aamodt, A. and E. Plaza, 1994. Case-based reasoning: Foundational issues, methodological variations and system approaches. AI Communications, 7(1): 39-59.
20. Pous, C., P. Gay, A. Pla, J. Brunet, J.Y. Sanz, T.R. Cajal and B. López, 2008. Modeling reuse on case-based reasoning with application to breast cancer diagnosis. In Artificial Intelligence: Methodology, Systems, and Applications (pp: 322-332). Springer Berlin Heidelberg.
21. Pal, S.K. and S.C. Shiu, 2004. Foundations of soft case-based reasoning (Vol. 8). Wiley. Com
22. Dalal, S., V. Athavale and K. Jindal, 2011. Case retrieval optimization of Case-based reasoning through Knowledge-intensive Similarity measures. International Journal of Computer Applications, 34(3).
23. Hui, D., 2009. An improving method of CBR retrieval based on self-organizing map. In Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on (1: 616-620). IEEE
24. Atanassov, A. and L. Antonov, 2012. Comparative Analysis of Case Based Reasoning Software Frameworks JCOLIBRI and myCBR. Journal of the University of Chemical Technology and Metallurgy, 47(1): 83-90.
25. Frank, A. and A. Asuncion, 2010. UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California. School of Information and Computer Science, pp: 213.
26. Lotfy Abdrabou, E.A.M. and A.M. Salem, 2010. A breast cancer classifier based on a combination of case-based reasoning and ontology approach. In Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multi-conference on (pp: 3-10). IEEE