Research on Optimization of Case-Based Reasoning System

Lin Tong
SOVO Entrepreneurial Center
Dalian Neusoft University of Information,
Dalian, China
e-mail: tongin@neusoft.edu.cn

Di Wu
School of Information Science & Engineering, Northeastern University
Shenyang, China
e-mail: wudi_ky@neusoft.edu.cn

***Abstract*—In this paper the deduction and optimization scheme is proposed for CBR decision making system. The CBR system is improved by using CURE_KNN algorithm. Cases in the library are clustered into some subsets, and the standard case library is constructed in a hierarchical manner. After the similarity between the target case and the central index point of each subset is computed, the nearest neighbor is used for retrieval in the nearest neighbor subset. The case library is maintained with the case addition and deletion strategy based on clustering. The performance of the CBR system is improved by the above multiple optimization strategy. At last the efficiency and availability of the proposed scheme is verified with system test.**

Keywords-CBR; optimization; clustering algorithm; case retrieval

I. INTRODUCTION

Cluster analysis (Clustering Analysis), also known as clustering is an efficient data analysis technique widely used in various fields, originated from the role of cognitive science memory in human reasoning activities, stories and lessons of the past to solve new problems currently encountered in the decision-making method to solve the problem. Compared with the traditional rule-based problem solving methods, CBR has easy access to knowledge to avoid the knowledge acquisition bottleneck problem in traditional knowledge systems with easy knowledge base maintenance and no experts intervention.

CBR process is made up of problem discription, case retrieval, case matching, case revise, case study. It works as follows: given a newly defined problem, retrieve the most similar case set in the library to the current problem; by analysis of these solved issues, the approximate solution is necessarily revised and adjusted in order to meet the new needs until the solution to the current new problems is found; then the new examples is constructed and added to the case library. All aspects of the reasoning process are based on cases.

In this paper, the improved clustering technology is applied in the the CBR decision-making system. During the setup stage of case library, the test result sets are clustered by the improved CURE_KNN algorithm to identify the central points and set up indexes of these subsets. In the following retrieval process, the distance between the target case and each center is compared to select the subset with the largest similarity. Subsequently deduction is carried out in this case subset. In the case library maintenance, the corresponding model is described with pseudo-code based on clustering addition and deletion strategy.

II. IMPROVED CLUSTERING ALGORITHM

CBR systems often use clustering techniques to deal with the case library. An algorithm model based on combination of improved CURE hierarchical algorithm and K-NN algorithm is proposed for cluster analysis of a large amount of data and irregular case set. Improvements are made on the weakness that the CURE algorithm randomly selects data set and the K-NN algorithm sets the center of clustering. Through a combination of the two algorithms, the clustering accuracy in CURE algorithm is improved and large data set clustering in K-NN algorithm is more efficient. The improved algorithm of clustering idea is as follows:

First the entire data set is needed to be decomposed, using a high efficiency random decimation algorithm RSA to select a small part from clustering space as a sample data, and then through CURE clustering algorithm the data subset is divided into a group of partition for which the local clustering based on the smallest average distance is carried out to identify the center of each cluster and setup index. It is worth noting that the the clustering here by CURE clustering algorithm to find the center point is not a separate data, but certain representative cases and the clustering central point here is actually a cluster of data. The form of clustering data on behalf of a class avoids the simplicity of data to represent class, greatly improving the accuracy of data collection center point to retrieve the matching operation. After the establishment of a data center point index, and then use the K-NN algorithm to classify the entire data collection based on CURE algorithm to establish the center point index, after a certain number of iterations to repeat the process to identify the largest similarity results as the final poly class results.

III. IMPROVED CLUSTERING ALGORITHM IN THE CBR DECISION-MAKING SYSTEM

Cases extraction is a key technology in the CBR decision-making system which has a direct impact on the efficiency and quality of the CBR reasoning. The quality of the extracted instances determines whether the re-usage and revision of case is easy. Commonly used models include

nearest neighbor model, induction model, knowledge and guidance law extraction model, a variety of extraction methods and a combination of these models. Application of CBR is facing many challenges, the specific performance in practical applications, the number of cases in the case library grows at a very fast speed along with time and the number of applications, and the content is constantly changing, making it necessary to divide a large case library into several small case libraries. Otherwise it will directly affect the the the CBR extraction efficiency and performance of the CBR. In order to improve the system efficiency and ensure the reasoning performance of the CBR system, the aforementioned improved clustering algorithm, CURE_KNN, is applied to the CBR system.

IV. MAIN IDEA OF THE CASE LIBRARY CLUSTERING

The objects of the case library clustering are cases and cluster analysis is based on the similarity between the cases. Each attribute of a case is looked as one dimension which has its upper and lower bounds. All of the attributes form a multi-dimensional space with each case representing a point. The similarity between the cases reflect the distance in the multidimensional space, a larger similarity means a shorter distance, thus the possibility that the two cases belonging to the same cluster is bigger.

The main idea for clustering in CBR case library system is as follows: first the entire case library is decomposed, select a small part of the cases in the clustering space using the high efficient randomly extracting algorithm RSA. Then by CURE clustering algorithm divide the subset of cases into a group of local clustering, each division is based on the minimum average distance to find the center of each cluster and setup index. Then use the K-NN algorithm to cluster the entire case set based on the index of these cluster centers and the threshold value T created by CURE algorithm. After a certain number of iterations, the largest average similarity is identified as the final clustering result.

V. CASE LIBRARY RETRIEVAL PROCESS DESCRIPTION

Based on the cluster analysis of case library and the establishment of the index of each clustering center point using the improved CURE_KNN algorithm, to retrieve the target instances (DCase), first determine which cluster the target case belongs to using the combination of index and nearest neighbor searching algorithm. After that, the nearest neighbor algorithm is used to retrieve the case that has a most similarity with the target case. Nearest neighbor algorithm is efficient for well-organized and indexed library. The retrieval process is shown in Figure 1.

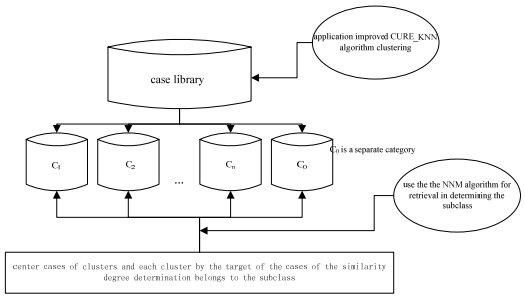

FIGURE 1. STRUCTURE OF CASE LIBRARY RETRIEVAL

The retrieval process includes the following stages:

First, determine a cluster that a target case belongs to. Calculate the similarity between target case and each center point case cluster. Identify the cluster that the center point case with the maximum similarity belongs to.

Second, compare the maximum similarity with the threshold T. If the maximum similarity is larger than the threshold T, put the target case into the cluster with the maximum similarity. Otherwise put this case into a separate class, C0.

Third, find the source case with maximum similarity using the nearest neighbor searching algorithm. Compute the similarity between the target case and all the cases in this subset. Identify the source case that has the maximum similarity with the target case as the candidate case to solve the target case.

The algorithm description for case retrieval process is given as follows.

SUB RetrieveInCB（CB）

I = MaxSimilarityOfCore （ DCase ， SubSetCaseCore[]）；

Find cluster I that has the maximum similarity with target case DCase；

IF the maximum similarity < threshold（T） THEN I=0 ‘0 as a separate class；

FOR C=StartInCB_I TO C <> “” Extract cases by nearest neighbor algorithm；

SameDegree（C，DCase） ‘Dcase is the target case；

NEXT C

RETURN the source case that has the maximum similarity with DCase；

END SUB

The time complexity of this algorithm is O(Len(CB_I)), where, Len(CB_I), is the number of cases in the cluster I.

VI. DESCRIPTION OF CASES DATABASE MAINTENANCE PROCESS

In the CBR system constantly problem-solving process, more and more new cases are stored in the case library, case library will become bigger and bigger. When the number of cases is over the upper bound, the cost of the retrieval will be very high. This phenomenon is known as the Marsh (swamping Problem). It is necessary to maintain the case library when the Marsh happens.

Case Database maintenance operations include the case library index (Indexing), the increase in cases (Addition) and examples of deletion (Deletion) operation. The purpose of maintaining the case library is to ensure the effectiveness, efficiency and accuracy of the CBR system, which is especially important when the CBR system is used to solve practical problems. This is because for any application system its knowledge about the content and structure will change over time. For example, if new cases id constantly added, invalid cases will be removed. How to ensure the performance of the CBR system under the premise of effective maintenance of the case library has become a new hot spot for CBR research.

When conduct cases addition and removal, it is necessarily to update the index of the case timely in two ways: one is the dynamic incremental update, the other is case library. re-indexing.

From the aforementioned assumptions based on case reasoning, similar questions have similar solutions, and similar cases solve similar problems. Therefore, before the deletion policy is given, the case covering set (CoverSet) and the reachable case set (ReachSet) are pre-defined.

For a given case set C = {c1, ... , cn} c ∈ C For each of the cases, there are:

The case covering set is CoverSet (c) = {c '∈ C | S (c, c') <RC};

the reachable case set is ReachSet (c) = {c '∈ C | S (c, c') <RR}.

Where, S (c, c ') is the similarity between the case c and c', RC and RR are the radius of CoverSet and ReachSet respectively.

The CoverSet of a case determines the space that a case can solve a problem. The larger the set is, the stories, the stronger the ability to deal with the problem is. Such cases should be kept.

The ReachSet of a case determines to what degree that the case can be replaced by the cases in it. The larger the set is, the more cases can be replaced by other cases. the deletion of such cases will not influence the ability to solve problems in CBR system.

The following discussions are on the cases addition and deletion:

A. case addition

The cases addition can be divided into the following phases:

First stage: the maximum capacity of the case subset is set as K, two thresholds asT1 and T2 and T1 <T2;

Second stage: identify each center index case cluster using improved CURE_KNN clustering algorithm on library CB;

Third stage: Compute the similarity between the target case and the central point of each case subset. If the similarity is greater than the threshold T1, the target case should belong to the subset If the similarities of all central points are smaller than the threshold T1, the target case belongs to a separate class (C0);

Forth stage: Compute the similarity between the target case and all the cases in the subset it belongs to.If the similarity between C and target case is greater than T2, compare the CoverSet of the target case and similar case C. If the CoverSet of the target case is greater than that of C, C is replaced by the target case and the case addition completes. IF the similarity between the target Case and any case is not greater than T2, proceed to the next step;

Fifth stage: If the number of cases in case subset is less than the maximum capacity K, add the target case and the case addition completes. Otherwise Re-cluster the class and divide it into several subsets, compute the similarity between the target case and newly added center, select the subset with the maximum similarity and add it to this class.

B. case deletion

Traditional deletion policies include: random deletion policy, the Minton delete strategy. Here a new deletion policy is proposed. Based on the idea of creating a central index of clustering, this paper proposes the deletion policy, and the basic idea is the cluster analysis of the case library clustering, making cases with big similarity in the same cluster and leaving cases with small similarity to different clusters. Meanwhile the outlier cases should be cheated as separate classes, because the classes are often the core cases and generally they should not be deleted.

The case deletion based on clustering strategy can be divided into the following phases:

First stage: initialize parameter RC, RR;

Second stage: re-cluster cases in the case library (In our system the improved CURE_KNN algorithm is used). After this is done, the case library C = {C0, C1, ..., CK}, where, Ci ($0 \le i \le k$) is a cluster;

Third stage: in each cluster (except of C0) obtain the case with ReachSet;

Forth stage: If there are multiple cases with ReachSet, consider the size of their CoverSet, and delete cases with a smaller CoverSet. Otherwise delete cases with biggest ReachSet. Compared with the traditional case deletion strategy, proposed strategy can optimize the system case to the maximum degree, improve the case retrieval efficiency and system performance.

VII. CONCLUSION

The paper describes optimization expert decision-making system based on case reasoning using improved clustering algorithm, including the hierarchical construction of the optimized case library, case representation and storage, case retrieval and matching, case modification and learning. Under the premise of quality guarantee cases extraction, the speed of cases extraction is greatly improved. Meanwhile the case library addition and deletion is simplified. More importantly an effective method to extract cases in the large case library is proposed to improve the traditional clustering methods.

Clustering on case library in this paper is based on the improved CURE_KNN algorithm with the following features:

(1) Retrieval and maintenance of large case library is particularly effective; ensure the retrieval efficiency and quality, overcome the nearest neighbor retrieval method for

its disadvantages of low efficiency on large-scale case database searching, and simplify the case library maintenance operations;

(2) The algorithm has a strong self-learning ability so that the system can make more accurate decisions;

(3) The algorithm is simple, high efficient since it can deal with non-spherical and sharp size changing case and the isolated points in case library.

Future work on the proposed system include

(1) The clustering efficiency needs to be improved since it takes a long time;

2) The case retrieval and the policy modification need to be further developed to better improve the search efficiency;

(3) The revision and learning of case needs to be further studied to guarantee the system with a stronger self-learning ability.

REFERENCES

[1] Chen Juan, Yang Ying. Based on case-based reasoning teaching cases knowledge management system design [J]. Computer and Information Technology, 2010, 18(3): 57-59.

[2] Kevin Vogts, Nigel Pope. Generating Compact Rough Cluster Descriptions Using an Evolutionary Algorithm [J], Lecture Notes in Computer Science, 2004, 3103: 1332-1333.

[3] Hichem Frigui. SyMP: An Efficient Clustering Approach to Identify Clusters of Arbitrary Shapes in Large Data Sets [C], In Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining (SICKDD), 2002, pages 507-512.

[4] Anantaporn Srisawat, Tanasanee Phienthrakul, Boonserm Kijsirikul. SV-kNNC: An Algorithm for Improving the Efficiency of k-Nearest Neighbor [J], Lecture Notes in Computer Science, 2006, 4099: 975-978.

[5] Xipeng Qiu, Lide Wu. Nearest Neighbor Discriminant Analysis [J], International Journal of Pattern Recognition and Artificial Intelligence, 2006, 20(8): 1245-1259.