Merging clustering for improved case retrieval stage of the case-based reasoning

Komal[1], Charu Pujara[2]

[1]Scholar, M. Tech, MRIU, Faridabad

[2]Associate Professor, Dept. of CSE, MRIU Faridabad

Abstract— **The case-based reasoning is one of the budding approach which facilitate the addicts to resolve the impending difficulty with facilitate of the presented problem solving familiarity. But it countenance the crisis of being sluggish down when case base enlargement with time. In this paper the deduction and optimization scheme is proposed for CBR pronouncement creation system. The CBR system is being enhanced by using clustering algorithm with k-NN algorithm. Cases in the collection are being huddled into several subsets, and the standard case library is constructed in a hierarchical manner. After the resemblance between the goal case and the innermost index point of each subset is computed, the nearest neighbor is used for reclamation in case retrieval phase for fastening case-based reasoning system**.

Keywords— **Case-based reasoning, Case retrieval, Clustering, Cure algorithm**.

I. Introduction

The case-based reasoning is a problem-solving loom that replicates the human being problem-solving conduct. In this approach, the trouble is being solved out on origin of precedent experiences gained from throughout the process of solving the problem in the earlier period. In case of multifaceted system, it is extremely complicated to devise the circumstances with domain rules. Other shortcoming is that the rules necessitate supplementary input information than is characteristically accessible, because of imperfect problem specifications or because the knowledge needed is simply not available at problem-solving time. But in case of CBR approach, if general knowledge is not sufficient because of too many exceptions, or when new solutions can be derived from old solutions supplementary without difficulty than from scuff, then on foundation of precedent experiences, the problem is being solved.

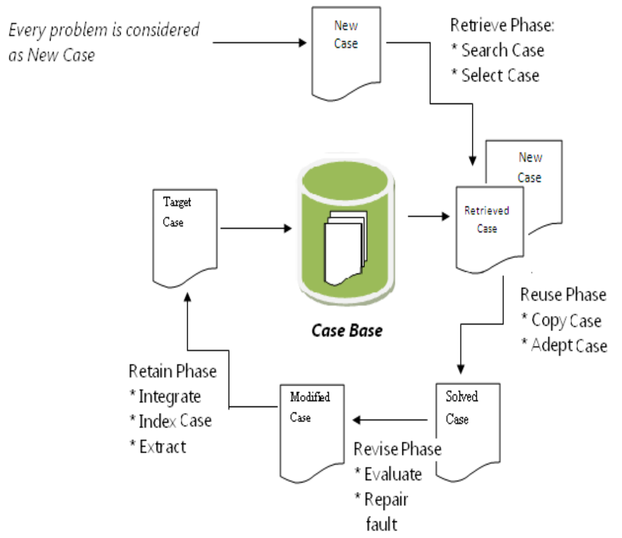

Figure 1: Case Based reasoning Cycle.

The case based reasoning involves four phases in the dilemma solving as given below:

- Retrieval phase.
- Reuse phase
- Revise phase
- Retain phase

Every problem pattern & its pledge are stored in form of the cases. It maintains the accumulation of the cases that is recognized as the case base. In this system, every problem is measured as the new-fangled case. In the retrieve phase according the new case, estimated solution case is being searched from the case base & preferred. After the assortment of the case, that case is adapted with the innovative case. It generates the resolved case. Now the solved case is evaluated in the revise phase & the faults in that case are being repaired. Now customized case is the resolution of the problem. This solution is stored in the case with appropriate index. This action is compulsory for extracting the cases very competently & rapid access to the cases in prospect.

Case-Based Reasoning is not constrained to the salvage of the experience. Another very important feature of case based reasoning is its coupling to learning. As the human beings learns from the precedent familiarity, the case base reasoning supports erudition from the history experience. Learning in CBR occurs as a usual by contraption of problem solving. When a problem is lucratively solved, the experience is retained in order to resolve comparable problems in the potential. When an attempt to solve a problem fails, the motive for failure is identified and remembered in order to shun the identical blunder in the future.

II. CASE RETRIEVAL PHASE

The Retrieve chore starts with a (fractional) problem explanation, and ends when a most excellent corresponding prior case has been found. Its subtasks are referred to as

- **Identify Features**: The identification task essentially comes up with a set of pertinent problem descriptors. To identify a problem may involve simply noticing its input descriptors, but often - and particularly for knowledge-intensive methods - a more elaborate approach is taken, in which an attempt is made to 'understand' the problem within its context. Unknown descriptors may be disregarded or requested to be explained by the user.
- **Initially Match**: the objective of the matching task is to revisit the situate of cases that are satisfactorily analogous to the new case - given a similarity threshold of some kind.
- **Search**: The task of finding a good match is typically split into two subtasks: An initial matching process which retrieves a set of plausible candidates, and a more elaborate process of selecting the best one among these. The latter is the Select task, described below. Finding a set of matching cases is done by using the problem descriptors (input features) as indexes to the case memory in a direct or indirect way. There are in principle three ways of retrieving a case or a set of cases: By following direct index pointers from problem features, by searching an index structure, or by searching in a model of general domain knowledge. The such cases are being searched
- **Select**: the selection task works on this set of cases and chooses the best match (or at least a first case to try out). From the set of similar cases, a best match is chosen. This may have been done during the initial match process, but more often a set of cases are returned from that task. The best matching case is usually determined by evaluating the degree of initial match more closely. This is done by an attempt to generate explanations to justify non-identical features, based on the knowledge in the semantic network. If a match turns out not to be strong enough, an attempt to find a better match by following difference links to closely related cases is made.

While several case-based approaches retrieve a prior case largely based on superficial, syntactical similarities among problem descriptors, some approaches attempt to retrieve cases based on features that have deeper, semantically similarities.

Table 1 Limitations of existing algorithms

Retrieval Techniques	**Strength**	**Weakness**
Nearest Neighbor Retrieval	*Simple*	Slow retrieval speed when the case base is large
Inductive Retrieval	*Fast retrieval speed*	1. Depends on pre-indexing which is a time-consuming process 2. Impossible to retrieval a case while case data is missing or unknown

A question that should be asked when choosing on a retrieval approach is the rationale of the retrieval task. If the purpose is to repossess a case which is to be adapted for reuse, this can be accounted for in the retrieval method. Approaches to 'retrieval for adaptation' have for example been suggested for retrieval of cases for design problem solving, and for analogy reasoning.

In some CBR tools, both techniques are used: inductive indexing is used to retrieve a set of matching cases, then nearest-neighbor is used to rank the cases in the set according to the similarity to the target case.

III. RELATED WORK

Yahia et al. (2011) combined fuzzy logic with case-based reasoning to identify useful cases that can support the DM. At the beginning, a fuzzy CBR based on both problems and actors' similarities was advanced to measure usefulness of past cases. For efficiency, they needed an optimal design of membership functions of fuzzy sets. Then, they relied on a meta-heuristic optimization technique i.e. Particle Swarm Optimization to adjust the parameters of the inputs and outputs fuzzy membership functions.

Bonissone et al. (2009) proposed a hybrid case-based reasoning (CBR) system for predicting the construction cost of high-rise buildings at the preliminary design stage. First, the extracted cost factors (CFs) of a high-rise building were shown to significantly improve the cost estimation system's performance. For developing a CBR system, a hybrid approach that combines CBR with genetic algorithms (GAs) for cost estimation was adopted. Genetic algorithms were used for optimized weight generation and applied to real project cases. Additionally, this paper proposes the identification of an alternative similarity score measurement formula. The proposed formula evaluated the contrast between the alternative case matching approach and the classical formula in a scenario involving the use of cost factors describing a case. The results indicated that the proposed GA-based CBR system could consistently reduce errors and potentially be useful to owners and contractors in the early financial planning stage. Accordingly, it was expected that the developed CBR system would provide decision-makers with accurate cost information to assess and compare multiple alternatives for obtaining the optimal solution and controlling the cost.

Sample et al. (2001) outlined an effective search algorithm for k-d trees that combined an optimal depth-first branch and bound (DFBB) strategy with a unique method for path ordering and pruning. This technique was developed for improving nearest neighbor (NN) search, but had also proven effective for k-NN and approximate k-NN queries. K-d trees had been widely studied, yet their complete advantages were often not realized due to ineffective search implementations and degrading performance in high dimensional spaces.

Aggour et al. (2004) designed the generic Case-Based Reasoning tool, implemented, and successfully used in two distinct applications. SOFT-CBR could be applied to a wide range of decision problems, independent of the underlying input case data and output decision space. The tool supplemented the traditional case base paradigm by incorporating Fuzzy Logic concepts in a flexible, extensible component-based architecture. An Evolutionary Algorithm had also been incorporated into SOFT-CBR to facilitate the optimization and maintenance of the system. SOFT-CBR relied on simple XML files for configuration, enabling its widespread use beyond the software development community. SOFT-CBR had been used in an automated insurance underwriting system and a gas turbine diagnosis system.

Paz et al. (2000) presented a system for automatically evaluating the interaction that exists between the atmosphere and the ocean's surface. Monitoring and evaluating the ocean's carbon exchange process was a function that requires working with a great amount of data: satellite images and in situ vessel's data. The system presented in this study focused on computational intelligence. The study presented an intelligent system based on the use of case-based reasoning (CBR) systems and offered a distributed model for such an interaction. Moreover, the system takes into account the fact that the working environment was dynamic and therefore it requires autonomous models that evolve over time. In order to resolve this problem, an intelligent environment had been developed, based on the use of CBR systems, which are capable of handling several goals, by constructing plans from the data obtained through satellite images and research vessels, acquiring knowledge and adapting to environmental changes. The artificial intelligence system had been successfully tested in the North Atlantic Ocean, and the results obtained would be presented in this study

Kang et al. (2014) argued and motivated that association analysis of stored cases could significantly strengthen SBR. They proposed a novel retrieval strategy USIMSCAR that substantially outperformed SBR by leveraging association knowledge, encoded via a certain form of association rules, in conjunction with similarity

knowledge. They also proposed a novel approach for extracting association knowledge from a given case base using various association rule mining techniques. They evaluated the significance of USIMSCAR in three application domains—medical diagnosis, IT service management, and product recommendation

IV. PERFORMENCE VALIDATION OF CLUSTERING ALGORITHMS

Clustering is analogous to classification in that data are assembled together. However, unlike classification, the groups are not predefined. Instead, the grouping is consummated by finding the relationships between data according to distinctiveness originated in the authentic data. The groups are called clusters. A term comparable to clustering is database segmentation, where like tuples (records) in a database are grouped mutually. Some basic features of clustering:

1. The figure of clusters is not known.
2. There may not be any a priori awareness concerning the clusters.
3. Cluster results are energetic.

Clustering algorithms themselves may be viewed as hierarchical or partitioned. With hierarchical clustering, a nested set of clusters is created. Each level in the hierarchy has a separate set of clusters. At the lowest level, each item is in its own unique cluster. At the highest level, all items belong to the same cluster. With the hierarchical clustering, the desired number of clusters is not input.

With partitional clustering, the algorithm creates only one set of clusters. These approaches use the desired number of clusters to drive how the final set is created.

Cure algorithm is both a hierarchical component and a partitioning component. First, a constant number of points, c, are chosen from each cluster. These well-scattered points are then shrunk toward the cluster's centroid by applying shrinkage factor, $\propto$. When $\propto$ is 1, all points are shrunk to just one-the centroid. These points represented the cluster better than a single point (such as a medoid or centroid) could. With multiple representative points, clusters of unusual shapes (not just a sphere) can be better represented. CURE then uses a hierarchical clustering algorithm. At each step in the agglomerative algorithm, clusters with closest pair of representative points are chosen to be merged. The distance between them is defined as the minimum distance between any pair of points in the representative sets from the two clusters.

V. RESULT

There are the deficiencies of the *k*-NN algorithm given below:

- **Defining precise Similarity Measures:** Saying that a database object is the "adjacent neighbor" of the uncertainty implies that we have a way to gauge distances between the inquiry and database objects. The way we prefer to compute distances can drastically influence the accurateness of the system. At the same time, defining a first-rate remoteness compute can be a difficult task. For example, what is the right way to measure similarity between two cases? A research problem that we are incredibly fascinated is designing methods for mechanically learning a distance quantify given many illustration of pairs of analogous objects and pairs of dissimilar objects.
- **Inefficient Retrieval:** As mentioned earlier, finding the adjacent neighbors of the inquiry can be protracted, chiefly when we have a huge catalog. The problem can be even inferior when the detachment measure we employ it computationally exclusive. At the equivalent time, computationally exclusive distance measures are often used in computer vision and pattern appreciation in general. Examples of such measures embrace the edit distance for strings, the chamfer distance and Hausdorff matching for edge images, the Kullback-Leibler distance and the Earth Mover's Distance for probability distributions, and bipartite matching for sets of features. As familiarity expands in many different domains and ever larger databases are used to store that knowledge, achieving proficient retrieval becomes increasingly important, and at the same time gradually more exigent.

With the help of projected algorithm we try to conquer the problem with presented case-based reasoning system. In this paper, we present an innovative algorithm called modified *k*-

Nearest Neighbors to competently search precise *k* nearest training objects for an inquiry objective.

Table 7.1 evaluation between K-NN & projected algorithm

Factor	k-NN algorithm	Proposed Algorithm	BIRCH Algo.
Stages	Single	Double	Double
Time complexity	$O(d*n^2)$	O(d*log(n))	O(d*n)
Efficiency	Only with diminutive dimension case-base	toil capably when case-base's dimension grow	Less efficient
Handling imprecise data	No	Yes	No
Training data availably	More data is required	Less training data is required	More data is required

This algorithm has two stages. In the buildup stage, we separate the dataset into clusters and record the distance from each training object to its closest cluster center. In the searching stage, we first calculate the distances from a query object *q* to all cluster centers. This proposed algorithm incorporates two simple methods, the *k*-means clustering and the triangle inequality, into the nearest neighbors searching and achieves good performance compared to other algorithms.

VI. CONCLUSION & FUTURE SCOPE

The paper illustrates optimization of the specialist decision-making organization based on case reasoning using superior clustering algorithm which comprises the hierarchical creation of the optimized case collection, case illustration and storage, case repossession and identical, case amendment and erudition.

Under the assertion of excellence certification in cases extraction, the momentum of cases extraction is significantly improved. in the meantime the case library addition and deletion is cut down. The supplementary significantly effectual technique is being proposed to dig up cases in the hefty case library to recover the conventional clustering methods' limitations.

In the future scope, the distributed CBR approach has been proposed as an appropriate substitute, and research efforts in this area have established a number of supplementary benefits for favoring this over the more conventional approach. For example, any centralized CBR system will have limited coverage characteristics, so by drawing on extra knowledge that may be available but scattered throughout the system, coverage can be improved as can the overall competence of the system. Secondly, the more obvious benefit in terms of performance is that efficiency can be improved by distributing the workload. Finally, there are also potential gains in terms of system maintenance since it may be easier to adapt local case bases independently of each other

REFERENCES

[1] J. Kolodner, "Case Based Reasoning. Morgan Kaufmann," San Francisco, 1993.

[2] S. W. Changchien, M. C. Lin, "Design and implementation of case based reasoning system for marketing plans," Expert Systems with Applications, vol. 28, pp. 43–53, 2005.

[3] J. Sebestyénová, "Case-based Reasoning in Agent-based Decision Support System", *Acta Polytechnica Hungarica*, Vol. 4, No. 1, 2007, pp.127-138.

[4] 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.

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

[6] 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.

[7] Yong-Bin Kang, Shonali Krishnaswamy, and Arkady Zaslavsky, "A Retrieval Strategy for Case-Based Reasoning Using Similarity and Association Knowledge", IEEE transactions on cybernetics, vol. 44, no. 4, april 2014 pp. 473-487.

[8] 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: pp.975-978.

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

[10] G. R. Beddoe and S. Petrovic, "Selecting and weighting features using a genetic algorithm in a case-based reasoning approach to personnel rostering," Eur. J. Oper. Res., vol. 175, no. 2, pp. 649–671, 2006.

[11] K. Bradley and B. Smyth, "Personalized information ordering: A case study in online recruitment," Knowl.-Based Syst., vol. 16, nos. 5–6, pp. 269–275, 2003.

[12] C. M. Vong, P. K. Wong, and W. F. Ip, "Case-based classification system with clustering for automotive engine spark ignition diagnosis," in Proc. 9th Int. Conf. Comput. Inf. Sci., Aug. 2010, pp. 17–22.

[13] F. Azuaje, W. Dubitzky, N. Black, and K. Adamson, "Discovering relevance knowledge in data: A growing cell structures approach," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 30, no. 3, pp. 448–460, Jun. 2000.

[14] Z. Y. Zhuang, L. Churilov, F. Burstein, and K. Sikaris, "Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners," Eur. J. Oper. Res., vol. 195, no. 3, pp. 662–675, 2009.