Chapter 88
DDR: A Multidimensional Case Retrieval Optimization Algorithm

Jing-Bin Wang and Xuan Hu

Abstract Ontology-based case retrieval system, the efficiency of case retrieval will increase as the number of cases continue to lower. In this paper, a multi-dimensional case retrieval optimization algorithm, the algorithm through the multi-dimensional case dimensionality reduction into clusters of two-dimensional space, using a two-dimensional spatial clustering to represent a collection of case, and this two-dimensional spatial clustering to establish the R-tree spatial index, by the two retrieval methods to the multidimensional case retrieval. Proved that the method not only improves the accuracy of case retrieval, but also greatly improve the efficiency of case retrieval.

Keywords Case retrieval · R-tree index · Relative point · Relative vector

88.1 Introduction

Case-Based Reasoning (CBR) is a solution of similar problems in the past with results to establish a case library, the new case obtain a similar solution cases in the case base to adapt to the current strategy, which is an important field of artificial intelligence reasoning method [1, 2]. System of case-based reasoning, case retrieval efficiency related to the efficiency of the entire system, is the case

J.-B. Wang · X. Hu (✉)
College of Mathematics and Computer Science, Fuzhou University,
Fuzhou 350001, Fujian, China
e-mail: 43806652@qq.com

J.-B. Wang
e-mail: wjbcc@263.net

W. Lu et al. (eds.), *Proceedings of the 2012 International Conference on Information Technology and Software Engineering*, Lecture Notes in Electrical Engineering 211,
DOI: 10.1007/978-3-642-34522-7_88, © Springer-Verlag Berlin Heidelberg 2013

retrieval efficiency of case retrieval results [3]. Growing system use case library will make the gradual reduction in the efficiency of case retrieval, this phenomenon is known as swamp phenomenon [4]. In every case in the previous case retrieval than similarity are demand once traversal, that every retrieval are the full match search queries gradually grow to a certain size will reduce the efficiency of this in the case base. Huan-tong [5, 6] in view of this situation, a clustering algorithm is applied to the maintenance of the case base. Zheng [7] proposed a K-Means clustering algorithm. Changzheng [8], the feature weighting C-means clustering algorithm (WF-C-means) and clustering-based case retrieval program to create the index, due to C-means clustering algorithm to adjust the weights of all attributes are included in the difference in the definition of the difference in the definition adopted by the retrieval and new cases similar case is very precise and objective. Li [9] proposed an improved k-means clustering case retrieval algorithm to solve the clustering error due to noise, to evaluate the ability of cases by collecting user feedback and Case energy force as the rules of the selected sample cases.

The paper proposes a multi-dimensional case optimization algorithm the DRR: First point to the case of multidimensional space in the case base through dimensionality reduction calculation, the two-dimensional spatial clustering, and re-established the two-dimensional space clustering R-tree index; fault by dimensionality reduction calculation, the current failure of the two-dimensional representation; through the R-tree index lookup can help us to quickly locate the current failure of the two-dimensional clustering where then KNN algorithm in two dimensional spatial clustering multidimensional case, find the closest current fault near multidimensional Case.

88.2 Multidimensional Case Retrieval Optimization: DRR Algorithm: DRR

88.2.1 Case Dimension Reduction

In case-based reasoning system, set the case library $CS = (c_1, c_2, c_3,\ldots c_n)$ a nonempty finite set composed by the n-th case, $\ni c_i\ (0 \le i \le n) c_i \in CS_0$

The case of CS is usually based on the feature vector representation. Can be set case $c = (f_1,f_2,\ldots f_i\ldots f_m)$ is a nonempty finite set, where $f_i(1 \le i \le m)$ is a feature item of c, characterized term is used to describe the case a property.

Definition 1 $\forall a,b\ a,b \in CS$ then the case space distance between two points for the Euler distance R_{ab}

$$R_{ab} = \sqrt{\sum (a_i - b_i)^2} (1 \le i \le m)$$

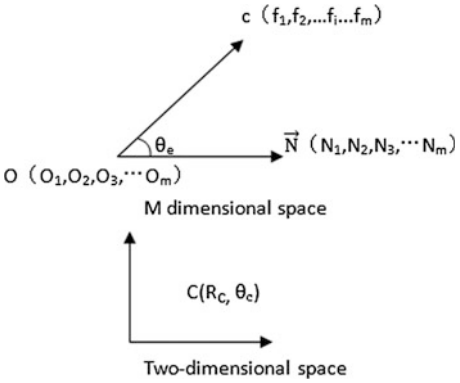

Fig. 88.1 Case spatial point dimensionality reduction

Definition 2 ∀a,b a,b ∈ CS, then the angle between the two vector case space for θ_{ab}.

$$\theta_{ab} = \cos^{-1} \frac{\sum a_i * b_i}{\|a\| * \|b\|} (1 \le i \le m)$$

Assuming an m-dimensional global reference point O (O_1,O_2,O_3,….O_m), a global reference vector $\vec{N}$(N_1,N_2,N_3,…N_m). According to definition 1 and definition 2, we calculate the relative distance of the case point c and O in the case space R_{co} (This simplified denoted as R_c), the angle between the case point c and vector $\vec{N}$ the reference angle θ_{co} (This simplified referred to as θ_c), then m dimensional case point c can be expressed in a two-dimensional spatial point C(R_c, θ_c), as shown in Fig. 88.1.

Space points in the m-dimensional case, there will be the same as the number of the global reference point O (O_1,O_2,O_3,…O_m) the relative distance, and with the global reference vector $\vec{N}$ (N_1,N_2,N_3,…N_m) having the same folder corner points, thus obtaining the case of a two dimensional clustering. We define:

Definition 3 Let a point D(R_d, θ_d) in a two-dimensional space S, represents a set of points (d_1,d_2,…d_i…d_n). d_i ∈ CS, the two-dimensional space of where d_i Represents both the D(R_d,θ_d).

Definition 4 Target case e(e_1,e_2,e_3,…e_m), the representation in the two-dimensional space is E(R_e, θ_e). ∀a(a_1,a_2,a_i…a_m) a ∈ CS; the representation in the two-dimensional space is A(R_a, θ_a). If E(R_e,θ_e) and A(R_a, θ_a) is closest in the ltwo-dimensional space S, then the Case point A is the most similar point of the target case e in the case space CS.

Proof According to Fig. 88.2, in the space CS in the m-dimensional case by the reference point O (O_1, O_2, O_3,….O_m), the fault case point e (e_1,e_2,e_3,…e_m), similar case point a(a_1,a_2,a_i…a_m) of a plane defined by three points forming a triangle △

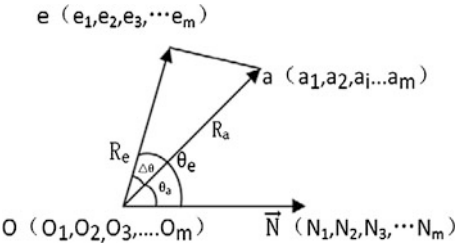

Fig. 88.2 The case points in the m-dimensional case base CS

OAE. Wherein R_e is the relative distance between the e and O (the length of the triangle sides θ_e), R_a is the relative distance between A and O (the length of the triangle sides θ_a), and the angle $\triangle\theta$ can be with the angle between a and the angle $\vec{N}$ between the e subtraction can be obtained, $\triangle\theta = |\theta_a - \theta_e|$.

In order to prove the case point a is similar to the fault case e that calculate the minimum Euler distance R_{ea} of the case point e with the case point a in the space CS (triangle edges ea's length).

By the law of cosines can be obtained in the case of m-dimensional space CS

$$R_{ea} = \sqrt{R_e^2 + R_a^2 - 2R_a R_e \cos\Delta\theta}$$

$$\Rightarrow R_{ea} = \sqrt{R_e^2 + R_a^2 - 2R_a R_e + 2R_a R_e - 2R_a R_e \cos\Delta\theta}$$

$$\Rightarrow R_{ea} = \sqrt{(R_a - R_e)^2 - 2R_a R_e(1 - \cos\Delta\theta)}$$

$$\Rightarrow \lim_{\substack{R_a \to R_e \\ \Delta\theta \to 0}} \sqrt{(R_a - R_e)^2 - 2R_a R_e(1 - \cos\Delta\theta)}$$

The derivation of the formula can be obtained If the R_e of the current fault e infinitely approaching the R_a of point a in case library, and the angle between the e and a infinitely close to 0 (θ_e infinitely close to θ_a), in the two-dimensional space representation of S is E(R_e, θ_e) and A(R_a, θ_a) two points infinitely close, then limit R_{ea} will infinitely close to 0, that is, to prove the highest degree of similarity between the e and a in the case base CS.

88.2.2 Indexing R-Tree

Definition 5 Point D(R_d, θ_d) is a point in a two-dimensional space S, the minimum bounding MBR of point D is M_d, wherein M_d to ($R_d - \triangle r$, $\theta_d - \triangle\theta$), ($R_d + \triangle r$, $\theta_d + \triangle\theta$) for the two pairs of corner points of the rectangular range.

Point D(R_d,θ_d) in the R-Tree index is represented as a leaf node D′(ID, M_d), where ID is the case point identification ID, M_d is a the minimum outsourcing rectangle of the point D, all the space point to the establishment of the R-tree index.

88.2.3 Case Initial Search

Definition 6 Current fault e (e_1,e_2,e_3…..e_m), expressed in two-dimensional space S E(R_e, θ_d), and the scope of the query point E rectangular the M_{se} based (R_e,$-\triangle sr$, $\theta_e - \triangle s\theta$), ($R_e + \triangle sr$, $\theta_e + \triangle s\theta$) for the two pairs of diagonal points, wherein $\triangle sr$, $\triangle s\theta$ as the point E in the scope of the query, the query higher rectangular greater the accuracy of the query. Set a two-dimensional space S in the case of point set D, represented as R tree M_d, M_d and M_{se} intersect, then the case point set D is a similar set of points of the point E in a two-dimensional space S, whereby the query intermediate nodes Result set R(R_1,R_2,R_3…R_n).

The assumption that the junction point of the R-tree index is T, a new look up on current fault e(e_1,e_2,e_3…..e_m), find the highest similarity to the case of the case e algorithm described as follows:

(1) The calculation of e with a global reference point O(O_1,O_2,O_3,…O_m) of the relative distance R_e, calculating the angle θ_d of the global reference vector $\vec{N}$ (N_1,N_2,N_3,…N_m) and case e in the two-dimensional, so the mapping point in the space S is E (Re, e).
(2) Calculation the query range rectangular M_{se} of E(R_e, θ_e) in the R-tree index.
(3) If T is not a leaf node then check M_{se} whether intersect with M_t intersect then recursively check E intersects T sub-node, if the disjoint abandon find the node.
(4) If the T is a leaf node, it is determined whether all records in the T intersects with the M_{se}. So that we can find the record of the intersection of all the E in R-tree, and find along the branches of the tree down to find, and not to traverse the tree in each record.
(5) To find all the intermediate results set R (R_1,R_2,R_3…R_n) M_{se} intersect.

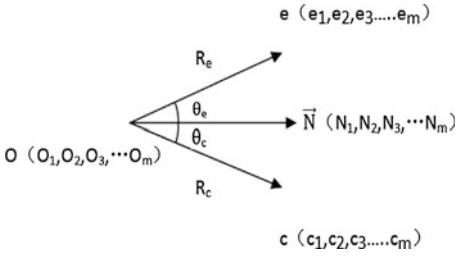

Fig. 88.3 Intermediate result set of case retrieval

88.2.4 Case Filter

Filter the intermediate result set R to calculate e (e_1,e_2,e_3…..e_m) and intermediate result concentrate all cases the similarity, and get the highest similarity case points.

The intermediate result set R may be the case in Fig. 88.3:

At this time, for the case c, the $R_e - \triangle sr >= R_c <= R_e + \triangle sr, \theta_e - \triangle s\theta >= \theta_e <= \theta_e + \triangle s\theta$, case c is a case point in the intermediate result set R, but the case c is not similar to the case of the fault case e, dimension reduction and lead to the concentration of the two-dimensional space point case point difference becomes smaller, thus redundant case point intermediate result set will be similar to the case c, when the needs of the middle result set to be filtered, and calculating e(e_1,e_2,e_3…..e_m) and intermediate result on the similarity of all cases, and the highest similarity to the case of point.

In this paper, the K-nearest neighbor (K-Nearest Neighbor Algorithm KNN) algorithm to calculate the similarity. It will be the case of feature vectors as points in a high-dimensional space, and then looking for a match with the current failure point in the problem space, will exceed the similarity threshold case is returned to the user, the general process is described as follows:

Input to be retrieved the fault case e output case object which similar to the case e. Results set R have k cases, each case by m attributes described, that $x_i = \{x_{i1},x_{i2},\ldots,x_{ij},\ldots,x_{im}\}$, i = 1,2,…, k; j = 1,2,…,m, calculated case x_i with the current fault e of between the Euclidean distance d(x_i,e):

$$d(x_i, e) = \sqrt{\sum \left(x_{ij} - e_j\right)^2} (1 \le j \le m)$$

Based on this, it is easy to know the similarity between the existing cases xi with the current fault case e:

$$SIM(x_i, e) = 1 - d(x_i, e)$$

88.2.5 Case Studies and R-Tree Index Correction

For new case e (e_1,e_2,e_3…..e_m), if the case base CS exists the case c (c_1,c_2, c_3…..c_m), and $c_i = e_i$ ($1 \le i \le m$), then say that the new case in the case library already exists, no new case with index correction.

If there is more than is required for the new case, Case Study:

(1) New case e is inserted into the case base CS.
(2) Calculation of the relative distance R_e of the new case e with the global reference point O, the angle θ_c between the case e and the global reference vector $\vec{N}$, the m-dimensional case point e can be expressed as two-dimensional spatial point E(R_e, θ_e).

(3) In two-dimensional space S for the new point E(R_e, θ_c), determine the point in space already exists, if there is illustrated in this cluster already exists, and is just a new clustering increasing identification ID; case point if there is no point in the space, a new clustering points, you need to re-adjust the R-tree index.

88.3 Experiment

We conducted a number of experiments in order to verify the validity of the DRR algorithm, the DRR algorithm and the traditional retrieval method based on rough set K-means algorithm [8] to compare the case base using a fault diagnosis system data, test case data were 30000, 50000, 80000, 100000, 200000, 300000, 500000. Algorithm by eclipses, JVM maximum memory 512 m.

88.3.1 Parameter Settings

The case spatial dimension m = 10, the reference point O (O_1,O_2,O_3,…O_m) is 0 (0,0,0,…, 0), the reference vector $\vec{N}$ (1,0,0,…0), the two-dimensional space point outsourcing rectangular size as $\triangle$ R = 1, $\triangle\theta$ = 1, the case similarity threshold value of 0.9, the query of the current failure rectangular range as $\triangle$ sr = 60 $\triangle$ sθ = 30.

88.3.2 Experimental Results and Analysis

88.3.2.1 Precision Experiments

The standard set of test cases has been given experiment (i.e query case base test case similarity is greater than the number of cases of 0.9), the query results closer to the standard result set is to illustrate the higher accuracy of the query. The experimental results shown in Table 88.1. As can be seen from Table 88.1, the K-means algorithm is more sensitive to the noise data when the data set of the query 50000, thereby affecting the accuracy caused by the clustering of the deviation due to the noise data. Noise data for the results of a query of the DRR algorithm is almost no effect, and the DRR algorithm result set high accuracy, stability to be higher than the K-means algorithm.

Table 88.1 Precision experiments

Case library size	Standard result set	K-means		DRR	
		Result	Accuracy (%)	Result	Accuracy (%)
30000	237	216	91.13	236	99.57
50000	410	308	75.12	405	98.78
80000	662	641	96.82	656	99.09
100000	829	807	97.34	817	98.55
200000	1630	1434	87.97	1606	98.52
300000	2585	2235	86.46	2547	98.64

Table 88.2 Comparison table of the efficiency of several algorithms

Case library size	Traditional algorithm (ms)	K-means (ms)	DRR (ms)
50000	29.33	15	11
80000	31.78	16	13
100000	37	16	15
200000	67.67	51.67	36.67
300000	93.33	73.25	53
500000	Memory overflow	Memory overflow	235

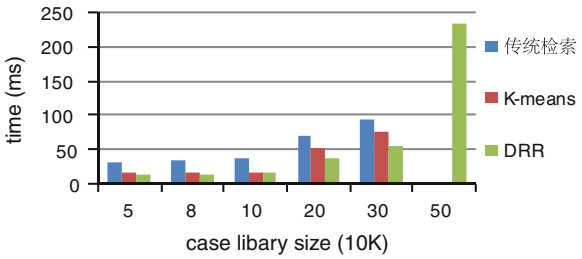

Fig. 88.4 Query results of the histogram

88.3.2.2 Query Efficiency Experiments

Our in 6 case data set of three algorithms query efficiency comparative experiments, the results as shown in Table 88.2 and Fig. 88.4.

Traditional retrieval methods every time all the data read memory query compared with the gradual growth of case space data, the time of the query data also gradually growth occurs when the data reaches a certain number of memory overflow. K-means and DRR algorithm to read data in the case base, indexing, each query are index-based query, so efficiency will be relatively fast indexing K-means based on the data of the entire case base, when the data reaches a certain amount of time will run out of memory; the DRR algorithm, dimensionality reduction means of two-dimensional clustering of the records in the case base, after for 2D clustering results establish the R-tree index, so that in memory R-tree

index space is much smaller than the index of the K-means algorithm. Datasets great when DRR algorithm can still achieve inquiries, not only does not appear out of memory and a good solution to the K-means algorithm search efficiency degradation caused by the sample points is too large, large data amount of case the retrieval efficiency of the library.

88.4 Conclusion

In this paper, a multi-case retrieval optimization algorithm DRR algorithm, the algorithm by clustering, two retrieval of dimensionality reduction method, not only speed up the retrieval efficiency, and the case of two-dimensional clustering is based on unrelated business, to avoid the classification errors caused by manual sorting, while avoiding the degradation caused due to the sample point is too large, the search efficiency, and to improve the case library retrieval efficiency of the large amount of data. The next step will algorithm to further improve weight impact on the clustering of feature items considering the case, in order to improve the accuracy and efficiency of case retrieval.

References

1. Aamodt A, Plaza E (1892) Case—based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications 1994 7(1):39–59
2. Clerk Maxwell J (1892) A treatise on electricity and magnetism, vol. 2, 3rd edn. Clarendon, Oxford, pp 68–73
3. Yin—Shan G, Qiang H, Yan Z et a1. (2003) Case-base maintenance based on representative selection for I-NN algorithm 2003 international conference on machine learning and cybemetics. pp 2421–2425 (in Chinese)
4. Derere L (2000) Case-based reasoning: diagnosis of faults in complex systems through reuse of experience. proceedings of international test conference, pp 77–105
5. Huantong G, Mingjun X, Xiang Z, Qingsheng C (2005) Research on application of clustering algorithm in CBM. Comp Eng 31(12):166–168
6. Yorozu Y, Hirano M, Oka K, Tagawa Y (1982) Electron spectroscopy studies on magneto-optical media and plastic substrate interface. IEEE Transl J Magn Jpn 2:740–741, Aug 1987. [Digests 9th annual conference magnetics Japan, p 301] (in Chinese)
7. Zheng F (2008) A rough-based k-means clustering algorithm. Computer engineering and applications 44(20):141–142 (in Chinese)
8. Changzheng L, Dong D (2010) Case indexing and retrieval based on clustering algorithm of weighted feature C-MEANS. Comp Appl Softw 27(2):111–114 (in Chinese)
9. Li Q, Huilin J (2011) Case retrieval algorithm based on k-means clustering. Comp Eng Appl 47(4):185–187 (in Chinese)