Abstract—The problem of optimizing distributed database includes: fragmentation and positioning data. Several different approaches and algorithms have been proposed to solve this problem. In this paper, we propose an algorithm that builds the initial equivalence relation based on the distance threshold. This approach differs from the previous works proposed by Hirano et al.,[10], [14] and Bean et al.,[2], [14-16] in that the proposed algorithm automatically determines the number of clusters based on the data set of survey.

In this paper, we use knowledge-oriented clustering techniques for vertical and horizontal fragmentation problem. The measure of similarity was developed based on from the available measure of the classical algorithms in data mining.

The paper is organized as follows: section 2 presents a brief overview of the basic concepts. We detail with the proposed vertical and horizontal fragmentation algorithms, in section 3 and section 4, respectively. We then discuss the main contributions of proposed approach in section 5.

I. INTRODUCTION

In distributed computing environments, each unit of data (item) which is accessed at the station, (site) is not usually a relationship but part of the relationship. Therefore, to optimize the performance of the query, the relations of global schema are fragmented into items.

There are several types of data fragmentation that are use vertical fragmentation, horizontal fragmentation, mixed fragmentation and derived fragments. Two classical algorithms associated with horizontal fragmentation and vertical fragmentation are PHORIZONTAL and BEA respectively [11]. Many authors have proposed solutions improved the above two algorithms, as Navathe, et al., (1984) [13], Chakravarthy, et al., (1994) [3],... However, the complexity of this algorithm is quite large, with vertical fragmentation problem is O(n^2), where n is the number of attributes and horizontal fragmentation is O(2^m), where m is the number of records [9],[11].

In recent years, several authors have incorporated to solve the problem of fragmentation and positioning, by using the optimal algorithms [5-6], [10] or using the heuristic method [4], [7]. The execution time of these algorithms is remarkably smaller than the classical algorithm.

The used technical association rules in data mining to vertical fragmentation has been mentioned in [8]. However, the data mining techniques do not attract many authors.

In this paper, we use knowledge-oriented clustering techniques for vertical and horizontal fragmentation problem. The used technical association rules in data mining to vertical fragmentation has been mentioned in [8]. However, the data mining techniques do not attract many authors.

In the clustering algorithm based on knowledge-oriented, we propose an algorithm that builds the initial equivalent relation based on the distance threshold. This approach differs from the previous works proposed by Hirano et al.,[10], [14] and Bean et al.,[2], [14-16] in that the proposed algorithm automatically determines the number of clusters based on the data set of survey.

The paper is organized as follows: section 2 presents a brief overview of the basic concepts. We detail with the proposed vertical and horizontal fragmentation algorithms, in section 3 and section 4, respectively. We then discuss the main contributions of proposed approach in section 5.

II. BASIC CONCEPTS

A. Vertical fragmentation

Vertical fragmentation is the collective decay properties of the relational schema R into the sub schema R_1, R_2, ..R_m, such that each attribute in these sub schemas is often accessed together.

To show how often the same queries together, Hoffer and Severance introduced the concept attribute affinity [11].

If Q = \{q_1, q_2, ..., q_m\} is a set of applications, then \(R(A_1, A_2, ..., A_d)\) is a relational schemas. The relationship between \(q_i\) and attributes \(A_j\) is determined by using the values:
use(qi, Aj) = If Aj is engaged in qi
0, A j is not engaged in qi

(1)

Put \( (A_i, A_j) = \{q \in Q \mid \text{use}(q, A_i), \text{use}(q, A_j) = 1\} \). Attribute affinity between \( A_i \) and \( A_j \) is:

\[
\text{Aff}(A_i, A_j) = \sum_{q \in Q} (\sum_{r \in \text{ref}(q)} \text{acc}(q))
\]

(2)

In particular, \( \text{refl}(q) \): the number of pairs of attributes \((A_i, A_j)\) is referenced in the application \( q \) at station \( S_i \); \( \text{accl}(q) \): frequency of access to applications \( q \) in station \( S_i \). BEA algorithm consists of two main phases:

1) Permutations row, column affinity matrix of attribute to obtain the cluster affinity matrix (CA) which has global affinity measure AM (global affinity measure) [11] is the largest.

2) Find the partition of the set of attributes from the matrix CA by exhaustive method, so that:

\[
Z = \text{CTQ} \ast \text{CBQ} - \text{COQ2} \text{is the maxima, with:}
\]

\[
\begin{align*}
\text{CTQ} &= \sum_{q \in \text{TQ} \wedge S_j} \sum_{r \in \text{ref}(q)} \text{acc}(q)
\text{CTQ} &= \sum_{q \in \text{TQ} \wedge S_j} \sum_{r \in \text{ref}(q)} \text{acc}(q)
\text{COQ} &= \sum_{q \in \text{OQ} \wedge S_j} \sum_{r \in \text{ref}(q)} \text{acc}(q)
\end{align*}
\]

| \( A_i \) | \( A_2 \) | \( \vdots \) | \( A_j \) | \( \vdots \) | \( A_n \) |
|---|---|---|---|---|---|
| \( A_i \) | \( TA \) | \( \vdots \) | \( \vdots \) | \( \vdots \) | \( \vdots \) |
| \( \vdots \) | \( \vdots \) | \( \vdots \) | \( \vdots \) | \( \vdots \) | \( \vdots \) |
| \( A_{i+1} \) | \( \vdots \) | \( \vdots \) | \( \vdots \) | \( \vdots \) | \( \vdots \) |
| \( A_{i+1} \) | \( \vdots \) | \( \vdots \) | \( \vdots \) | \( \vdots \) | \( \vdots \) |
| \( A_n \) | \( BA \) | \( \vdots \) | \( \vdots \) | \( \vdots \) | \( \vdots \) |

In which,

\[
\text{AQ}(q_i) = \{A_j \mid \text{use}(q_i, A_j) = 1\};
\]

\[
\text{TQ} = \{q \mid \text{AQ}(q_i) \subseteq TA\};
\]

\[
\text{BQ} = \{q_i \mid \text{AQ}(q_i) \subseteq BA\};
\]

\[
\text{OQ} = \text{Q} \backslash \{TQ \cup BQ\}
\]

The complexity of the algorithm is proportional to \( n^2 \).

### B. Horizontal Fragmentation

Horizontal fragmentation divides set records into a smaller set of records. Horizontal fragmentation is based on the query conditions, which are expressed through simple predicates of the form: \( A_i \theta <\text{value}> \).

Let \( Pr = \{Pr_1, Pr_2, \ldots, Pr_k\} \) is a set of simple predicates extracted from a set of applications. A conjunction of the predicates, which is built from \( Pr_i \) will have the form:

\[
P_1 \land P_2 \land \ldots \land P_n
\]

(3)

Where \( p_i^* \) is a predicate, which received one of \( p_i \) or \( \neg p_i \) values.

PHORIZONTAL algorithm uses the conjunction of the predicates \( P_1 \land P_2 \land \ldots \land P_n \) to find the conditions for horizontal fragmentation of data [9]. The relation \( r(R) \) will be fragmented into \( \{r_1(R), r_2(R), \ldots, r_k(R)\} \), with \( r_i(R) = \sigma \text{Pr}(r(R)) \), \( 1 \leq i \leq k \); \( F_i \) is a predicate which forms the conjunction of the primary predicates [9].

### C. Information systems and the inability to distinguish relationship

- The information system is a pair of \( SI = (U, A) \), where \( U \) is a finite set of objects \( \{t_1, t_2, \ldots, t_n\} \). \( A \) is non-empty finite set of attributes.

- An equivalent relation (A binary relations satisfy properties reflective, symmetric and transitive) defined on \( U \) is called an inability to distinguish relationship (irrespectively relationship) on \( U \).

### D. Clustering algorithm Knowledge-Oriented

Clustering algorithm Knowledge-Oriented based on rough set theory was first proposed by the authors Shoji Hirano, et al., [10], [15]. This is a clustering algorithm automatically determines the number of clusters based on the survey data [12]. The main idea of this clustering algorithm consists of 2 phases:

1) Created of equivalence relation on the set of object clustering.

2) Editing of the equivalence relation using a threshold Tk based on the measure irrespective. This iterative process will update \( Tk \) a best clustering results is obtained Using this algorithm to data fragmentation, we have proposed initial equivalence relation based on the average distance between objects.

Clustering algorithm Based on knowledge orientation, so we propose as follows:

**Input:** \( U \) = the set of objects to be clustered.

(Each object must be describe the information needed to construct a similar measure).
**Output:** The clusters (corresponding to the fragment of data).

**Method:**

**Step 1:** Construct a matrix of similarities \( S = S(t_i, t_j) \) between all pairs of objects \((t_i, t_j)\);

**Step 2:** Specify an initial ability to distinguish relationship \( R_i \) for each object. Synthesis to get an initial clustering;

**Step 3:** Construct ability to distinguish matrix \( \Gamma = \gamma(t_i, t_j) \) to assess the quality of clustering;

**Step 4:** Modify the clusters by the inability to distinguish relationship \( R_i^{mod} \) for each object to achieve the revised clustering;

**Step 5:** Repeat steps 3 and 4 until a stable clustering is obtained.

The inability to distinguish relationship corresponds to the \( j \)-th attribute:

\[
R_i = \{ (t_i, t_j) \in U \times U : d(t_i, t_j) \leq T_k ] \text{, with } j = 1, 2, \ldots, n \}.
\]

Where \( d(t_i, t_j) \) is the distance between two participants clustering.

The threshold \( T_k \) is determined as follows:

\[
T_k = \left[ \frac{\sum_{j=1, j \neq i}^{n} (1 - s(t_i, t_j))}{(n-1)} \right] (n-1)
\]

With \( s(t_i, t_j) \) is the similarity measure of two objects \( t_i, t_j \).

**III. VERTICAL FRAGMENTATION PROBLEM BASED ON KNOWLEDGE-ORIENTED CLUSTERING TECHNIQUES**

Vertical fragmentation problem is converted to the clustering problem, based on the following concepts:

**A. Attribute and the reference feature vector**

**Definition 1:** The reference measure of transaction \( q_i \) on attribute \( A_j \), denoted by \( M(q_i, A_j) \):

\[
M_{ij} = M(q_i, A_j) = \text{use}(q_i, A_j) \cdot f_i
\]

In which \( M_{ij} \) is the frequency with which transactions \( q_i \) refer to attribute \( A_j \). With \( f_i \) is the frequency of transactions \( q_i \) and use \( q_i, A_j \) is defined by formula (1).

**Definition 2:** \( VA_j \) reference feature vector of attribute \( A_j \) with reference transactions \( \{ q_1, q_2, \ldots, q_m \} \) is defined as follows:

\[
VA_j = [M_{ij}, M_{2j}, \ldots, M_{mj}]
\]

**B. The similarity measure of two properties**

**Definition 3:** The similarity measure of two attributes \( A_k \), \( A_l \) has two feature vectors corresponding to the reference transactions \( \{ q_1, q_2, \ldots, q_m \} \):

\[
VA_k = (M_{1k}, M_{2k}, \ldots, M_{mk})
\]

\[
VA_l = (M_{1l}, M_{2l}, \ldots, M_{ml})
\]

Is determined by the cosine measure:

\[
s(A_k, A_l) = \frac{\sum_{i=1}^{m} M_{ik} \cdot M_{il}}{\sqrt{\sum_{i=1}^{m} M_{ik}^2 \cdot \sum_{i=1}^{m} M_{il}^2}} \quad (5)
\]

**C. Vertical fragmentation based on knowledge-oriented clustering technique**

To illustrate the vertical fragmentation algorithm based on knowledge-oriented clustering techniques. We use the assumption of examples about vertical fragmentation problem based on BEA algorithm is presented in [1, 8]:

The set of attributes \( A = \{ A_1, A_2, A_3, A_4 \} \)

The set of transactions \( Q = \{ q_1, q_2, q_3, q_4 \} \). The matrix used:

\[
\begin{array}{cccc}
A_1 & A_2 & A_3 & A_4 \\
1 & 0 & 1 & 0 \\
0 & 1 & 1 & 0 \\
0 & 1 & 0 & 1 \\
0 & 0 & 1 & 1 \\
\end{array}
\]

The frequency of application execution with a set of transactions \( \{ q_1, q_2, q_3, q_4 \} \), and \( F = \{ f_1, f_2, f_3, f_4 \} = \{ 45, 5, 75, 3 \} \).

From the assumption, we have the reference feature vectors:

\[
\begin{array}{cccc}
VA_1 & VA_2 & VA_3 & VA_4 \\
45 & 0 & 0 & 0 \\
0 & 5 & 75 & 0 \\
45 & 5 & 0 & 3 \\
0 & 0 & 75 & 3 \\
\end{array}
\]

The similar matrix \( S_{VA} = (s(A_k, A_l)) \) \( k=1,4; \ l=1,4 \)

\[
\begin{array}{cccc}
A_1 & A_2 & A_3 & A_4 \\
1 & 0 & 0.9918 & 0 \\
1 & 0.0073 & 0.9970 & 0 \\
1 & 0.0026 & 0 & 1 \\
1 & & & \\
\end{array}
\]

The result of the vertical fragmentation algorithm based on the clustering algorithm towards knowledge-oriented.

| Cluster | Set of attributes |
|---------|------------------|
| 1       | \{A_1, A_3\}    |
| 2       | \{A_2, A_4\}    |

This fragmentation results correlate with the results of the vertical fragmentation by algorithm BEA.
IV. **Horizontal fragmentation problem based on knowledge-oriented clustering technique**

Similar to vertical fragmentation, assuming conversion of horizontal fragmentation problem from PHORIZONTAL algorithm is based on the concept of the following establishments:

A. **Vectorization records of a relationship**

Considering the relations \( r(R) = \{T_1, T_2, \ldots, T_l\} \), the set of simple predicates extracted from applications on \( r(R) \) is \( Pr = \{Pr_1, Pr_2, \ldots, Pr_m\} \). Vector binary of records by the rule:

| \( Pr_1 \) | \( P_{i2} \) | \( \ldots \) | \( P_{ij} \) | \( \ldots \) | \( P_{im} \) |
|---|---|---|---|---|---|
| \( T_1 \) | \( a_{i1} \) | \( a_{i2} \) | \( \ldots \) | \( a_{ij} \) | \( \ldots \) | \( a_{im} \) |
| \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) |
| \( T_i \) | \( a_{i1} \) | \( a_{i2} \) | \( \ldots \) | \( a_{ij} \) | \( \ldots \) | \( a_{im} \) |
| \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) |
| \( T_j \) | \( a_{j1} \) | \( a_{j2} \) | \( \ldots \) | \( a_{jj} \) | \( \ldots \) | \( a_{jm} \) |

\[ \forall a_{ij} = \begin{cases} 1, & \text{if } T_i \{Pr_j\} = \text{true} \\ 0, & \text{if } T_i \{Pr_j\} = \text{false} \end{cases} \]

B. **The similarity measure of two binary vector**

Consider two vectors \( x_i \) and \( x_j \), that are represented by binary variables. Assuming binary variables have the same weight. We have event tables as Table 3. Where \( q \) is the number of binary variables equal to 1 for the two vectors \( x_i \) and \( x_j \), \( s \) is the number of binary variables equal to 0 for simpler \( x_i \) but equal to 1 for \( x_j \), \( r \) is the number of binary variables equal to 1 for \( x_i \) but is 0 for \( x_j \), \( t \) is the number of binary variables equal to 0 for all vectors \( x_i \) and \( x_j \).

| Object \( j \) | \( 1 \) | \( 0 \) | Sum |
|---|---|---|---|
| Object \( i \) | \( q \) | \( r \) | \( q+r \) |
| \( 0 \) | \( s \) | \( t \) | \( s+t \) |
| Sum | \( q+s \) | \( r+t \) | \( p \) |

- The difference of two vectors \( x_i \) and \( x_j \) based on the symmetric binary dissimilarity are:

\[ d(x_i, x_j) = \frac{r + s}{q + r + s + t} \]

- The similarity measure between two vectors \( x_i \) and \( x_j \) is defined by the Jaccard coefficient:

\[ \text{sim}(x_i, x_j) = 1 - d(x_i, x_j) \]

C. **Horizontal fragmentation based on knowledge-oriented clustering techniques**

Considering the relations \( EMP \), [11]:

| ENo | EName | Title |
|---|---|---|
| T1 | E1 | Jioe |
| T2 | E2 | M.Smith |
| T3 | E3 | A.Lee |
| T4 | E4 | J.Smith |
| T5 | E5 | B.Casey |
| T6 | E6 | L.Chu |
| T7 | E7 | R.David |
| T8 | E8 | J.Jone |

Consider two simple predicates:

\[ p_1 = \{\text{Title} > "Programmer"\}; \quad p_2 = \{\text{Title} < "Programmer"\}, \]

with string comparison rules in alphabetical order. Vectorization records by two predicates \( p_1 \) and \( p_2 \) are:

| \( p_1 \) | \( p_2 \) |
|---|---|
| T1 | 1 | 0 |
| T2 | 0 | 1 |
| T3 | 1 | 0 |
| T4 | 0 | 0 |
| T5 | 0 | 1 |
| T6 | 1 | 0 |
| T7 | 1 | 0 |
| T8 | 0 | 1 |

D. **The result horizontal fragmentation of the relation (EMP)**

The result horizontal fragmentation of the relation EMP as Table 3 based on the clustering algorithm towards knowledge-oriented. We have used the similarity measure defined by formula (7), where \( d(x_i, x_j) \) is calculated by the formula (6).

With \( k=2 \), we have:

| Cluster | The set of records |
|---|---|
| 1 | T1, T2, T6, T7 |
| 2 | T2, T4, T5, T8 |

With \( k=3 \), we have:

| Cluster | The set of records |
|---|---|
| 1 | T1, T3, T5, T7 |
| 2 | T2, T5, T8 |
| 3 | T4 |
And with k=4, we have:

| Cluster | The set of records |
|---------|--------------------|
| 1       | T₁, T₃            |
| 2       | T₂, T₃, T₄        |
| 3       | T₅                |
| 4       | T₆, T₇            |

This fragmentation results coincide with the results of the horizontal fragmentation by algorithm PHORIZONTAL.

V. Conclusion

In this paper, we used knowledge-oriented clustering techniques for fragmentation problem in distributed database systems. With this solution, we also proposed transforming hypothetical of this problem to hypothetical of clustering problems. Experimental results on the data in [11] that results correlate with the results obtained from the classical fragmentation algorithms PHORIZONTAL and BEA. In addition to experimental data as presented, we also tested on a number of different data set. The results are also similar to the two classical algorithms above. In the future work, we will carry out the Analysis of large data sets that to compare test the usability of the proposed solution.

VI. References

[1] Adrian Runceanu (2007), Towards Vertical Fragmentation in Distributed Databases, International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering, CISSE 2007.

[2] C.L Bean, C.Kambhampati (2008), Autonomous Clustering Using Rough Set Theory, Inter-national Journal of Automation and Computing, Vol.5 (No.1). pp. 90-102. ISSN 1476-8186.

[3] S. Chakravarthy, J. Muthura, R. Varadarajan, and S. Navathe (1994), An objective function for vertically partitioning relations in distributed databases and its analysis, Tech Rep. UF-CIS-TR-92-045.

[4] Hui Ma, Klaus-Dieter Schewe and Markus Kirschberg (2006), A Heuristic Approach to Vertical Fragmentation, Proceedings of the conference on Databases and Information Systems IV: Selected Papers from the Seventh International Baltic Conference DB&IS’2006.

[5] I. Lungu, T. Vatluu, A. G. Fodor (2006), Fragmentation solutions used in the projection of Distributed Database System, Proceedings of the 6th International Conference "ELEKTRO 2006", Edis-Zilina University Publishers.

[6] Jiawei Han, Micheline Kambel (2012), Data Mining: Concepts and Techniques, 3rd ed Mor-gan Kaufmann Publishers.

[7] Eng. Marwa F. F. Areed, Prof.Dr.Ali EI-Dosouky, Hesham A. Ali (2008), A Heuristic Approach for Horizontal Fragmentation and Allocation in DOODB, INFOS2008, March 27-29, 2008 Cairo-Egypt © 2008 Faculty of Computers & Information-Cairo University.

[8] Narasimhaiah Gorla, Pang Wing Yan Betty (2010), Vertical Fragmentation in Databases Using Data-Mining Technique, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

[9] Shahidul Islam Khan, A. S. M. Latiful Hoque (2010), A New Technique for Database Fragmentation in Distributed Systems, International Journal of Computer Applications (0975 – 8887), Volume 5—No.9.

[10] Shoji Hirano and Shusaku Tsumoto (2001), A Knowledge-Oriented Clustering Technique Based on Rough Sets, Computer Software and Applications Conference, COMPSAC’01.

[11] Tamer O., ValduriezP (1999), Principles of Distributed Database Systems. Prentice Hall Englewood Cliffs, Second Edition, New Jersey 07632.

[12] Yin-Fu Huang, Jyh-her Chen (2001), Fragment Allocation in Distributed Database Design, Journal of Information Science and Engineering, pp. 491-506.

[13] Navathe S, Ceri S, Wiederhold G, Dou J (1984), Vertical partitioning algorithms for database design, ACM Trans Database Syst 9(4).

[14] Hoffer, J.A., And Severance, D.G (1975), The use of cluster analysis in physical database design. In Proceedings 1st International Conference on Very Large Databases (Framingham, Mass).

[15] A. S. Darabant (2005), A New Approach In Fragmentation Of Distributed Object Oriented Databases Using Clustering Techniques, Studia Univ. Babes»{Bolyai, Informatica, Volume L, Number 2.

[16] A.S. Darabant, A. Campan & others (2004), AI Clustering Techniques: A New Approach in Horizontal Fragmentation of Classes with Complex Attributes and Methods in Object Oriented Databases, Proc. of the Int. Conf. on The-ory and Applications of Mathematics and Informatics.

Personal information:

- Full name: Luong Van Nghia, year of birth 1964.
- Address: 99 Hung Vuong street, Quang Ngai city, Quang Ngai province, Vietnam.
- Academic degree: Master of Computer Science
- Name of Agency: Faculty of Informatics, Pham Van Dong Universisty
- Phone:: (+84) 913 498 804, email: nghia.itq@gmail.com (email2: lvnghia@pdu.edu.vn)
- Graduation year: Master in Computer Science, Hue University of Science
- Research Areas: Data mining, Distributed Database, Embedded System.