WKTD: A Novel Algorithm for Reducing Search Time Using Data Mining Mechanism

Arash Ghorbannia Delavar[1], Somayeh Zare Harofteh[2], Majid Feizollahi[3], Nasim Anisi[4]

[1 2 3] Payame Noor University,Tehran, Iran
[4] Payame Noor University, OIIC Company , Tehran, Iran

Summary
In this paper we will present an A Novel Algorithm for Reducing Search Time Using Data Mining Mechanism. This method is achieved when we create a threshold detector (TD) by which we may conduct clustering so that in data base accumulation, we may present a new competence function by portioning max and min point which produce specific intervals in the information accumulation. In WKTD algorithm, by evaluating parameters used in data collection in the data base of Iranian Workers Association and Oil Industry Investment Company, the recommended algorithm can be used for searching the abovementioned data bases with a high record of estimated costs which are obtained from data collection. Delays considered in searching banks also have been assessed and sweep time and return time of task search have been calculated using competence function. In order to reduce repetitive data in the data base, we were able to present a new method using the threshold detector which enables us to create repetitive data by clustering. Compared with basic K-Means and WKMSD, the recommended algorithm has a higher performance and dependability and is more reliable than previous algorithms.

Key words:
WKTD, ERPSD, ERPASD

1. Introduction

In the twentieth century, environmental conditions for entering customers are among powerful managerial tools which are used in information technology. For processing cumulative percentage of information, we need to do high level data investigations at data bases. This can be integrated with data mining technique so that we may achieve accomplishments for reducing data gauging. In environmental conditions, frameworks have been used in which there are algorithms as rules that might create a process in shaping customers' satisfaction level so that this structure may make the system more efficient. But if integrated frameworks and algorithms cannot enhance the level of customers' satisfaction, they produce parameters the use of which not only do not increase system's productivity but also decreases its dependability. But if the same parameters are integrated with inter-related framework and algorithm or, in other words, if the recommended algorithm is properly incorporated into the framework, it can enhance the satisfaction level of the customers. Now, by observing systems distributed by data mining mechanism, we use a series of algorithms used in basic K-Means and WKMSD as well as ERPSD framework. In this method, by integrating combined methods a suitable strategy has been presented which enhances the level of clients' satisfaction. Therefore, in order to increase the level of clients' satisfaction, we have replaced integrated central systems distributed integrated systems so that their dependability and security will be improved By studying previously Proposed frameworks and models, we succeeded in presenting suitable strategies for creating the new algorithm. Considering the new Proposed algorithm, we devised new technical methods in which a new methodology has been used because although all the above-mentioned independent algorithms have some advantages but after combing effective factors, their performance decreases. Using the Proposed WKTD algorithm, we succeeded in improving the performance of the base K-Means algorithm and WKMSD algorithm.

2. Distributed systems with data mining mechanism

Successful implementation of ERP systems with regard to their capabilities has always been studied and in this matter the most effective factors are selection of suitable process and algorithm.[11] Knowledge discovery process, extracting suitable information, and using optimal algorithms are among the most important challenges in this regard. Taking into consideration their unique requirements, various companies have presented different strategies for data extraction and knowledge discovery.[12] ERPSD is a new method for integrating all current processes and optimizing the employment of system resources with regard to competitive level and clients' satisfaction using ERP. In this method, security and performance levels have increased compared with the previous model and, thus, time and cost have decreased. Acquiring the new methodology requires identification of ERPDS phases and data mining algorithms.[2] Following the review of the previous literature, it was found out that this methodology is better implemented in distributed rather than central systems. Therefore, in order to increase

Manuscript received March 5, 2011
Manuscript revised March 20, 2011

the efficiency, we use dependable distributed integrated systems so that we can remove problems observed in central ERPs.[8,9]
In a previous study, the ERPASD algorithm has increased security and performance levels as comparison with previous methods and also considering this algorithm, in the distributed data base, a technical method has been presented which reduces the total cost of the distributed integrated system, calculates repetitive data using Apriori ERPSD, and optimizes the ERP system. [1]
In order to control the data base and accessibility, there are various algorithms for data mining such as: K-Means, Page Rank, EM, and SVM Apriori. [3] But in this research, following an accurate and comprehensive study of various algorithms, we used K-Means algorithm which produces practical efficient results.

Table 1. Users Respond Time and cost In K-Means Algorithm(s)

		number of records	Users Respond Time In K-Means Algorithm(s)
1	S1	5000	0.022
2	S2	7000	0.033
3	S3	10000	0.037
4	S4	20000	0.039
5	S5	30000	0.041
6	S6	50000	0.043
7	S7	70000	0.048
8	S8	100000	0.049
9	S9	200000	0.254
10	S10	300000	0.277
11	S11	400000	0.436
12	S12	500000	0.491
13	S13	700000	0.691
14	S14	1000000	0.745
15	S15	2000000	1.155
16	S16	3000000	1.367
17	S17	4000000	1.915
18	S18	5000000	2.276
19	S19	6000000	2.781
20	S20	7000000	2.944
21	S21	8000000	3.257
22	S22	9000000	3.663
23	S23	10000000	3.891

3. Base K-Means algorithm

Base K-Means algorithm is one of the preferable data mining algorithms in which clustering method is used. In a simple type of this method, first, random points equal to required clusters are selected. Then the data are related to one of these clusters according to their proximity to these clusters and new clusters are formed. [3]Repeating the same procedure, in each repetition by averaging the data new centers can be estimated for them and again the data can be related to new clusters. This procedure continues until no change occurs in the data.

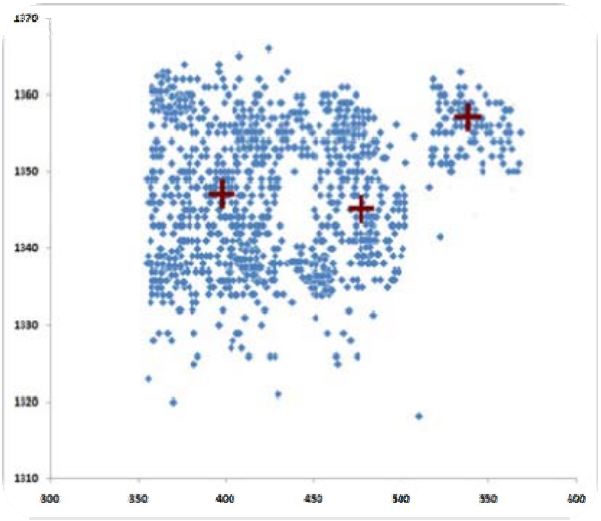

Fig 1. Function of K-Means algorithm in the data base of members of Khane Kargar center

The following function is the target function:

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2$$

Number of fields: Fc
Number of clusters: K
K = Fc + 1

$\| xi(j) - cj \|2$ is a criterion for the distance between xi(j) data points and cj cluster centre and j is the distance between n data points from their respective cluster centers. The problem with implementing this algorithm is that by determining the points and clustering in voluminous data, still data abundance accumulation percentage in each cluster will be high and survey time in data base will be higher compared with similar algorithms and we must present an algorithm which may reduce its search time.

4. WKTD Proposed algorithm

Using WKTD algorithm, we create a new technique by which data accumulative percentage at central points will be reduced. In this case, time survey in the data base and accumulation of information is presented by breaking up the areas. We will illustrate these phases by defining all the parameters affecting the WKTD algorithm.
WKTD algorithm has a higher performance that basic K-Means and WKMSD, but the important point here is that

information accumulation for large and small data is equally broken and this creates the problem of repetitive data in accumulation frequency percentage. We overcame this problem by using a new technique in WKTD. Since it is different from basic K-Means and WKMSD algorithms, we present a threshold detector (TD) in the algorithm which may balance the data accumulation. The threshold detector usually helps us manage the gauging time of the information in the data base and eventually reduce it. This is done when we make the breakage point selection using the threshold detector.

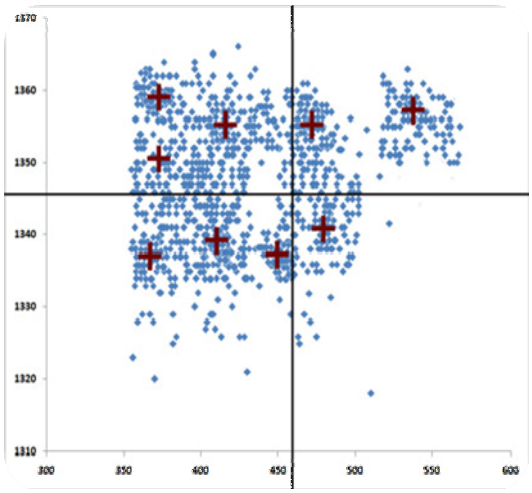

Fig 2. Function of WKTD algorithm in the part of data base of members of workers house

In WKTD algorithm a technique has been developed so that using the competence function and threshold detector, we may use the information accumulation percentage using the clustering method and create a suitable method for quick processing. This is achieved when we optimize distance, time, and cost using some parameters.
The following function is the target function:

$$j = \sum_{\frac{M(j)+m(j)}{2}}^{k} \sum_{\frac{M(i)+m(i)}{2}}^{n} \left\| \chi_{\frac{M(i)+m(i)}{2}}^{\frac{M(j)+m(j)}{2}} - C_{\frac{M(j)+m(j)}{2}} \right\|^2$$

Table 2. Hardware and software Information Used In Simulation

	Hardware or Software	Information
1	Processor	Intel 1.86Hz
2	Memory(RAM)	2GB
3	Operation System	Microsoft windows XP professional version 2003 SP3
4	System Type	32-bit operating system

M=Maximum , m=Mininmum
Fc =Number of fields
Rc=Number of records
Number of records in any section after average calculation
Rcc=
TD= Number of clusters in any section according to Data abundance accumulation
K=Number of clusters
RND=Round
Rcm=Minimum records for clustering

$$TD = RND\ \left(\frac{R_{cc} * F_c * (F_c + 1)}{R_c}\right)$$

$$K = \sum_{1}^{F_c^2} TD$$

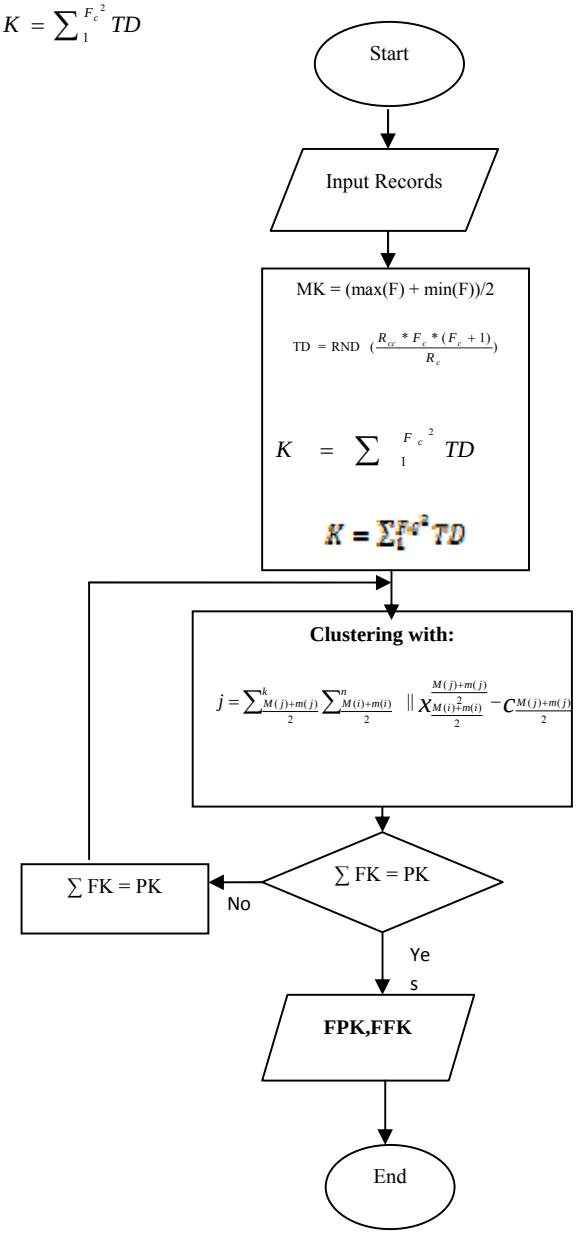

Fig 3. WKTD Flowchart

Table 3. Users Respond Time and cost In WKTD Algorithm(s)

		number of records	Users Respond Time In WKTD Algorithm(s)
1	S1	5000	0.017
2	S2	7000	0.024
3	S3	10000	0.028
4	S4	20000	0.029
5	S5	30000	0.031
6	S6	50000	0.033
7	S7	70000	0.036
8	S8	100000	0.041
9	S9	200000	0.141
10	S10	300000	0.229
11	S11	400000	0.374
12	S12	500000	0.438
13	S13	700000	0.577
14	S14	1000000	0.689
15	S15	2000000	0.861
16	S16	3000000	1.283
17	S17	4000000	1.427
18	S18	5000000	2.031
19	S19	6000000	2.319
20	S20	7000000	2.571
21	S21	8000000	2.846
22	S22	9000000	3.082
23	S23	10000000	3.476

$\| X_{\frac{M(j)+m(j)}{2}}^{\frac{M(j)+m(j)}{2}} - C_{\frac{M(j)+m(j)}{2}}^{\frac{M(j)+m(j)}{2}} \|^2$ is a criterion for the distance between $X_{\frac{M(j)+m(j)}{2}}^{\frac{M(j)+m(j)}{2}}$ data points and $C_{\frac{M(j)+m(j)}{2}}^{\frac{M(j)+m(j)}{2}}$ cluster centre and j is the distance between n data points from their respective cluster centers.

TD is a variable which calculates the number of clusters in each part based on the accumulation of data frequency.

In this method, when the information amount is high data abundance is broken in each area and survey time increases.

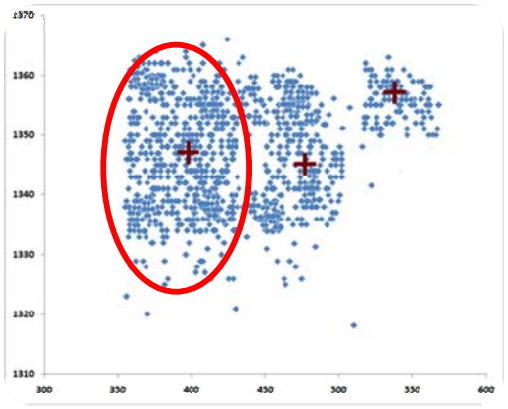

Fig 4. Data abundance accumulation percentage in K- Means final clustering

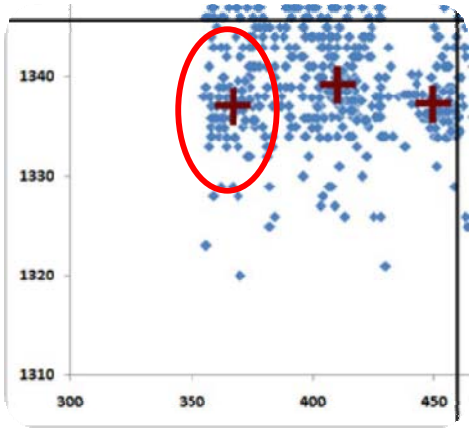

Fig 5. Data abundance accumulation percentage in WKTD final clustering

5. Conclusion

Taking into consideration the limitations which Base K-Means algorithm has caused in the data abundance accumulation percentage, we succeeded in reducing this problem by presenting a new technique with WKTD algorithm and, using this algorithm with a threshold detector, we created a target function in order to improve the survey method in the data base. Several simulations which have been studied using simulation software and data base as well as comparisons made indicate that the Proposed algorithm has proved its efficiency in comparison with Base K-Means algorithm and point finding survey in the Proposed algorithm for data accumulation has been performed with higher accuracy and a lower amount of survey. This algorithm has a higher efficiency compared with its previous similar algorithm and is more reliable than Base K-Means algorithm.

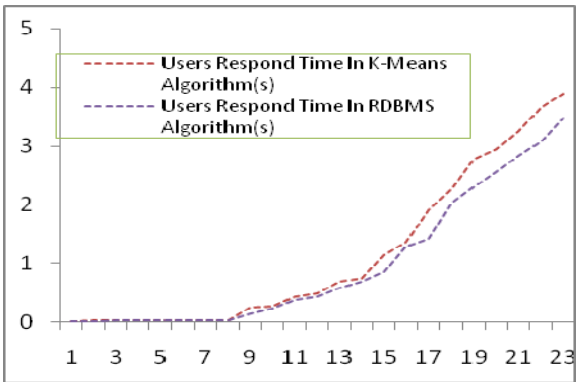

Fig 6. Result Respond Times of Implementation Algorithm In OIIC Data base

REFERENCES

[1] E.Bendoly "Theory and support for process frameworks of knowledge discovery and data mining from ERP systems" * Decision and Information Analysis, Room 411, 1300 Clifton Road, Goizueta Business School,Emory University, Atlanta, GA 30322-2701, USA Received 8 September 2001; received in revised form 12 February 2002; accepted 17 August 2002

[2] A. Ghorbannia Delavar, M. Zekriyapanah Gashti" ERPASD: A Novel Algorithm for Integrated Distributed Reliable Systems Using Data Mining Mechanisms" September 2010,Chongqing china

[3] A.Ghorbannia Delavar , V.Aghazarian , S.Sadighi "ERPSD: A New Model for Developing Distributed,Secure, and Dependable Organizational Softwares" IEEE CSIT 2009 Yerevan, Armenia, September 28 2009 ,PP.404-408.

[4] X.Wu , V.Kumar , J. R.Quinlan ,J.Ghosh , Q.Yang , H.Motoda, G.J. McLachlan , A.Ng · Bing Liu · Philip S. Yu , Z.H.Zhou , M.Steinbach , D.J.Hand "Top 10 algorithms in data mining" Dan Steinberg Received: 9 July 2007 / Revised: 28 September 2007 / Accepted: 8 October 2007 Published online: 4 December 2007© Springer-Verlag London Limited 2007

[5] A.Ghorbannia Delavar ,L.Saiedmoshir "Analysis of a New Model for Developing Concurrent Organizational Behaviors with Distributed, Secure and Dependable ERPSD Technology" IEEE CSIT 2007 September 24-28 2007,PP.321-324.

[6] A. Ghorbannia Delavar and M. Nejadkheirallah and M. Motalleb, A New Scheduling Algorithm for Dynamic Task and Fault Tolerant in Heterogeneous Grid Systems Using Genetic Algorithm, Chengdu, China, 2010.

[7] A.Abdullah , S.Al-Mudimigh , B.Zahid Ullah , C.Farrukh Saleem "A Framework of an Automated Data Mining Systems Using ERP Model" International Journal of Computer and Electrical Engineering (IJCEE 2009) , Vol. 1, No. 5 December 2009 , PP.651-655.

[8] Ch.Liu , Jiancheng An " Fast Mining and Updating Frequent Itemsets " International Colloquium on Computing, Communication, Control, and Management , IEEE CCCM 2008.

[9] W.Jin, Y.Wang, Y.Zhou, H.Wang " Research on Distributive Algorithm of Data Mining with Association Rules " 2009 IEEE

[10] K.Jin ,"A new algorithm for discovering association rules" ,2010 IEEE

[11] T.Lan, Catherine, Huang, D.Erdogmus , " A Comparison of Temporal Windowing Schemes for Single-trial ERP Detection " Proceedings of the 4th International IEEE EMBS Conference on Neural Engineering Antalya, Turkey, April 29 - May 2, 2009

[12] K.B. Hendricks a,1, V.R.Singhal b,*, J.K.Stratman b,2a R.Ivey "The impact of enterprise systems on corporate performance:A study of ERP, SCM, and CRM system implementations" School of Business, The University of Western Ontario, London, Ont., Canada N6A-3K7b College of Management, Georgia Institute of Technology, 800 West Peachtree St., NW, Atlanta, GA 30332-0520, United StatesAvailable online 23 March 2006

[13] N.A. Morton, Q.Hu "Implications of the fit between organizational structure and ERP:A structural contingency theory perspective" Department of Information Technology and Operations Management, Barry Kaye College of Business,Florida Atlantic University, Boca Raton, FL 33431-2007, USA

[14] C.Berchet a, G.Habchi b "The implementation and deployment of an ERP system:An industrial case study" ,*a Alcatel AVTF, 98 Avenue de Brogny, BP 2069, 74009 Annecy Cedex, France b LISTIC-ESIA, Domaine Universitaire, BP 806, 74016 Annecy Cedex, FranceReceived 29 March 2004; received in revised form 14 December 2004; accepted 13 February 2005 Available online 15 June 2005

[15] Neil A. Morton, Qing Hu]Implications of the fit between organizational structure and ERP:A structural ontingency theory perspective _Department of Information Technology and Operations Management, Barry Kaye College of Business,Florida Atlantic University, Boca Raton, FL 33431-0991, USA International Journal of Information Management 28 (2008) 391–402

[16] Elliot Bendoly Theory and support for process frameworks of knowledge discovery and data mining from ERP systems Decision and Information Analysis, Room 411, 1300 Clifton Road, Goizueta Business School, Emory University, Atlanta, GA 30322-2701, USA Received 8 September 2001; received in revised form 12 February 2002; accepted 17 August 2002

[17] Michel Pouly, Jean-Charles Binggeli, Rémy Glardon Ecole Polytechnique Fédérale de Lausanne (EPFL),ERP data sharing framework using the Generic Product Model (GPM) Souleiman Naciri, Naoufel Cheikhrouhou *, Laboratory for Production Management and Processes, Station 9, 1015 Lausanne, Switzerland 2010

[18] Arash Ghorbannia Delavar, Majid Feizollahi, Nasim Anisi "Using Knowledge Base for Work Flows of Distributed Integrated Systems by a proposed Algorithm with Data Mining Mechanism" January 26-28, 2011, Guiyang, China Paper ID : W25

[19] Arash Ghorbannia Delavar, Nasim Anisi and Majid Feizollahi," A New Framework for the Work Flows of Distributed Integrated Systems by Assessment of Effective Factors "March 11 - 13, 2011, Shanghai, China D0075

Arash Ghorbannia Delavar received the MSc and Ph.D. degrees in computer engineering from Sciences and Research University, Tehran, IRAN, in 2002 and 2007. He obtained the top student award in Ph.D. course. He is currently an assistant professor in the Department of Computer Science, Payam Noor University, Tehran, IRAN. He is also the Director of Virtual University and Multimedia Training Department of Payam Noor University in IRAN. Dr. Arash Ghorbannia Delavar is currently editor of many computer science journals in IRAN. His research interests are in the areas of computer networks, microprocessors, data mining, Information Technology, and E-Learning.

Somayeh Zare Harofteh
Birth date: 1980
Education: IT
Payame Noor University
Tehran, Iran

Majid Feizollahi
Birth date: 1978
Education: IT Management
Payame Noor University
Jame elmi karbordi university
(Khane karghar centre)
Tehran, Iran

Nasim Anisi
Birth date: 1982
Education: IT Management
Payame Noor University
OIIC Company
Tehran, Iran