International Journal on
"Technical and Physical Problems of Engineering"
(IJTPE)
Published by International Organization of IOTPE

ISSN 2077-3528
IJTPE Journal
www.iotpe.com
ijtpe@iotpe.com

June 2021 | Issue 47 | Volume 13 | Number 2 | Pages 144-148

APPLYING WORMHOLE APPROACH TO DESIGN A HIERARCHY IN A RELATIONAL DATABASE FOR QUICK DATA ACCESS

B. Ture Savadkoohi P. Nik Mohammadi

Department of Computer and Electrical Engineering, Seraj Higher Education Institute, Tabriz, Iran,
bita.turesavadkoohi@gmail.com, parviznik@gmail.com

Abstract- Databases are an important and impartible part of any organization in the modern era of information technology. Moreover, relational databases which are using for storing, retrieving and analyzing data are facing quick access to data when the amount of data is increasing. On the other hand, in theoretical physics field, wormholes are portals in space and time that are creating shortcut between two points which are far away from each others in one universe or two points from two different universes. In this paper, for having quick access to data, we aggregate the multi-layer graph model with relational model. Then, we use wormhole theory for creating shortcut between different nodes, e.g., far and near nodes. In order to prove the correctness of our method, we are applied K-Nearest Neighbors (KNN) algorithm for finding the shortcut between nodes. So that, for the purpose method of this paper, cosine distance gives best result among other functions.

Keywords: Relational Data Base, Wormhole, Shortcut, K-Nearest Neighbors.

1. INTRODUCTION

Relational data base is designed and implemented based on relational model concepts [1-3]. Quick access is the major problems of this kind of data bases when the amounts of data are increasing.

In the world of information and communication technology, wormhole attacks are occurred on wireless networks [4-5]. This type of attack is carried out by malicious nodes that create a shortcut between two nodes that change the path from source to destination. So that, data packets will be received by the far node.

Wormhole is based on General Relativity (GR) is used for creating shortcut to connect the distinct universes or distant regions in one universe [6-7], as shown in Figure 1.

On the other hand, data mining in brief is the study of collection, cleaning, processing, analyzing and converting raw data into useful information [8]. In the modern science research, data mining is a terminology for generating some forms of data either for diagnostic or analysis purposes that are encountered in real applications. Classification is a very important step in mining science: It is collection and separation of the same and different objects or entities [8-9].

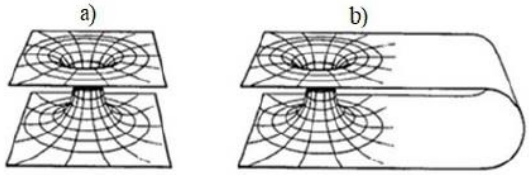

Figure 1. The geometric shape of standard wormhole. (a) connect two different universes (b) connect parallel universe [6]

The fast access to data in relational data base, seeks to answer following questions:

- How to reduce unnecessary surveying between nodes.
- How to prevent redundancy.

For this aim, the multi-layer graph model with relational model is aggregated. Then, the shortcuts between nodes are created.

The sequence of presentation of the paper is organized in the following manner: In Section 2 preliminaries about the structure of the exploited tables are presented. Related works are summarized in Section 3, while Section 4. provides creation of shortcuts in relational data base. The evaluation is illustrated in Section 5. Finally, the paper is concluded in Section 6.

2. PRELIMINARIES

As shown in Figure 2, the main components of the tables which are used in this paper are Parent-ID, Category-ID and Path. So that, Parent-ID of each record indicatives each node has been created by which parent and the value of this field is unique. Thus, in this way, the children of parents are found with respect to Parent-ID. Moreover, the value of this filed in the first record, due to its root property is exclusively zero. Since each parent can have several children, thus it can be repeated based on the number of children.

Category-ID indicatives the node number. In order to distinguish the nodes from each other, there is a need to assign unique number to each node. For this aim, Category-ID is chosen as sequential incremental type. Thus, in the case of parent's children are created sequential, the value of Category-ID of them are

sequential or close to each other. Otherwise, the value of Category-ID is different when the parent's children are created at different period of time.

Path-ID, determines the node's type. The node is real, if the value of Category-ID is equal to Path-ID, otherwise that node is shortcut and it is defined as outlier data.

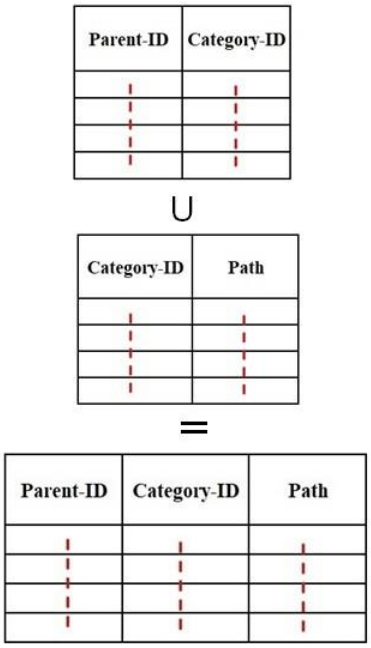

Figure 2. The structure of tables

3. RELATED WORKS

Classification is very important step of data mining science. There are several classification algorithms. Some of them are applied as follows:

The KNN classifier is one of the popular data mining method based on closest training example in the feature space [10-13]. Guha et al. [14] used clustering based on hierarchy in order to construct the hierarchical relationship among data. Shen et al. [15] described DBSACN algorithm for real-time image superpixel segmentation. Barati et al. [16] applied Fuzzy approach to detect Faulty reading to improve the decision-making algorithm. Wan et al. [17] introduced Neural-network classifier while Pernkopf et al. [18] described Bayesian network classifiers for optimization problem.

4. CREATING THE SHORTCUT IN HIERARCHY MODEL

In order to prove the correctness of proposed method, the KNN algorithm is performed by RapidMiner [10-13]. For this aim, a hierarchical structure is created in the relational model. Each record in the data base tables is considered as a vertex. So that, each vertex is included three fundamental dimensions such as: Parent-ID, Category-ID and Path. In the formation of the node, Parent-ID, Category-ID and Path are equally important. Although, the nodes in our design have two modes, either real or shortcut. Since, the values of the fields may change by normalizing the value of the Parent-ID, thus, the rule for determining the shortcut will be violated. For solving this problem, we consider the table as a two-layer graph such as core and shell layer (Figure 3).

Suppose V be set of Vertexes, RV be Real Vertex, SV be Shortcut Vertex, ES be the Edge of Shortcut, ESh be set of Edges of the Shell layer and ECo be set of Edges of the Core layer. So that, the union of two ESh and ECo sets represents the total set of edges. In the shell layer when there is self-loop, that vertex is RV. This means, the value of Category-ID and Path are equal. Otherwise, that vertex is SV. In this paper, SV is used for creating shortcut between near or far vertex. In fact, SV and ES plays the role of a transmitter to VR. Since all the fields have the same type, so normalization is not needed in applying the KNN algorithm. Moreover, the values of Category-ID and Path will keep and the shortcut for quick access to data will determine.

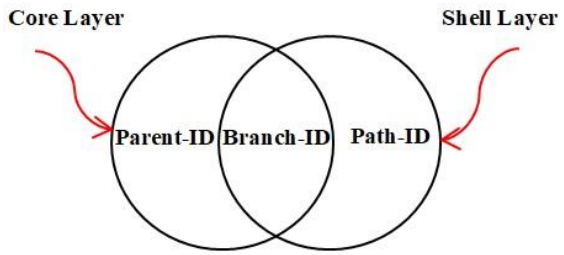

Figure 3. Illustration of core and shell layer

4.1. Calcification with KNN Algorithm

KNN is a machine learning algorithm which is used widely for both pattern recognition and classification applications [10-13]. The key component of this well-known non-parametric approach is calculating distance/similarity between the test samples and all training ones.

As shown in Figure 4, filled squares and empty circles are used as two classes of data [19]. The ⊕st data point will be classified as empty classes, in the case of applying 1-nearest neighbors. Otherwise, it will be classified as filled squares in the case of applying 3-nearest neighbors. Although, the class cannot be decided in the case of 2-nearest neighbors.

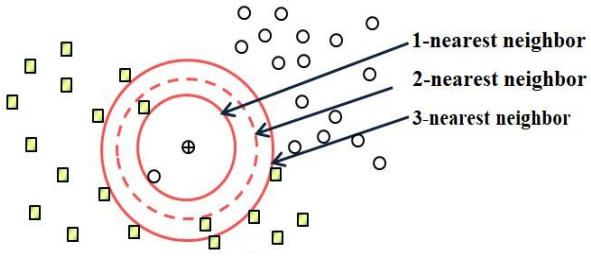

Figure 4. An example of KNN classification [19]

The main steps of KNN are:

- Step 1: Choose the number of nearest region (k).
- Step 2: Compute the distance/similarity between all training records and new objects.
- Step 3: Sort the data based on distances/similarity value in ascending order.

- Step 4: Estimate data by using these *k* distances.
- Step 5: Find in the *k* training records nearest to the object which are occurring most frequently.

The common used distance and similarity functions such as Euclidean distance, cosine similarity, cosine distance are defined as follows [9]:

$$\text{Euclidean Distance} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (1)$$

$$\text{Cosines similarity} = \frac{x.y}{\|x\|_2 \|y\|_2} = \frac{\sum_{i=1}^{d} x_i y_i}{\sum_{i=1}^{d} (x_i)^2 \sum_{i=1}^{d} (y_i)^2} \quad (2)$$

$$\text{Cosine distance} = 1 - \cos(\alpha) = \frac{x_i^T x_j}{\|x_i\| \|x_j\|} \quad (3)$$

where, x_i and y_i stand for the component of two records, respectively.

5. EVALUATION

In order to determine the performance parameters for analysis, the confusion matrix [20-22] that is illustrated in Table 1 is applied. So that, True Positive (TP) is the number of candidates of class that are already shortcut vertexes, and also begin classified as shortcut vertexes. The True Negative (TN) is the number of candidates which are real vertexes and also being classified as real vertexes. However, False Negative (FN) is the number of actual shortcut vertex candidates that are incorrectly begin classified as real vertexes. The False Positive (FP) is the numbers of real vertexes that are incorrectly begin classified as shortcut vertexes.

Table 1. Confusion matrix

	Recommended Item by the System	Item not Recommended by the System
Expected Item	TP	FN
Not an Expected Item	FP	TN

Accuracy, precision, recall, Fisher score (F-score) and specificity is calculated in order to evaluate the proposed method by Equations (4), (5), (6), (7) and 8, respectively [20-22]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (7)$$

$$\text{F-score} = \frac{2TP}{2TP + FP + FN} \quad (8)$$

Figures 5-9 depict the value of the accuracy, precision, recall, specificity and F-score that are calculated with 47, 60 and 77 outlier numbers.

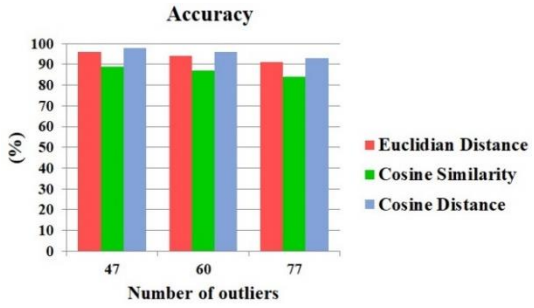

Figure 5. Accuracy measures with different distance and similarity functions

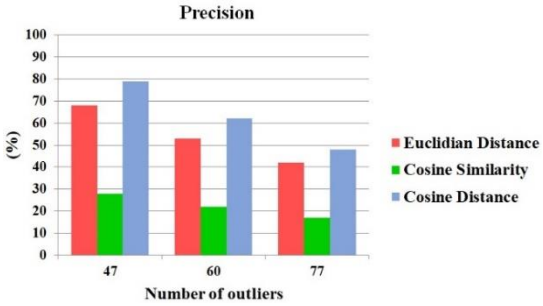

Figure 6. Precision measures with different distance and similarity functions

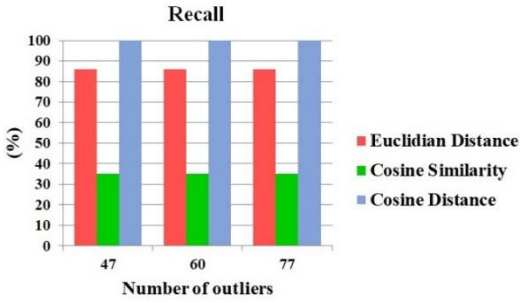

Figure 7. Recall measures with different distance and similarity functions

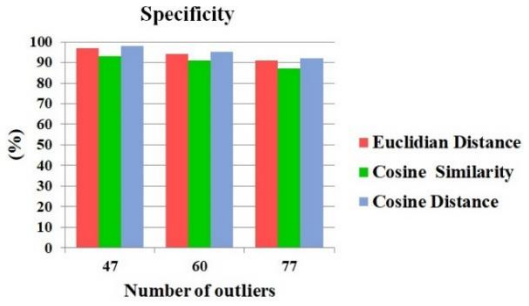

Figure 8. Specificity measures with different distance & similarity functions

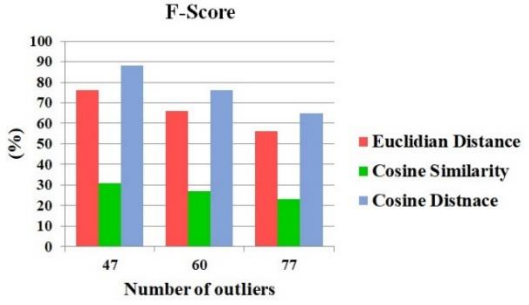

Figure 9. F-Score measures with different distance and similarity functions

6. CONCLUSIONS

Nowadays, relational data bases by increasing data are facing the problem of quick access to data. For this aim, we have created a hierarchical model in relational data base. So that, by applying wormhole approach the distances between nodes are reduced and the problem of redundancy in this model is solved. So that, there isn't any need to add similar nodes to the hierarchy, when the node is created in one branch and there is a need for that node in other branches, but it will create shortcut to that branch that node is there. The shortcut nodes have the address of the main node and referred to real node. For this aim, in order to proof the correctness of the proposed method of this paper K-Nearest Neighbors (KNN) is applied. Then, different distances and similarity functions was compared by calculating accuracy, precision, recall, specificity and F-score with different outlier numbers. The result shows the cosine distance is best function for proposed method of this paper.

REFERENCES

[1] R. Kraleva, V. Kralev, N. Sinyagina, P. Koprinkova Hristova, N. Bocheva, "Design and Analysis of a Relational Database for Behavioral Experiments Data Processing", International Journal of Biomedical Engineering, Vol. 14, No. 2, pp. 117-132, 2018.

[2] B. Grad, "Relational Database Management Systems: The Business Explosion", Journal of IEEE Annals of the History of Computing, Vol. 35, No. 2, pp. 8-9, 2013.

[3] N. Mallig, "A Relational Database for Bibliometric Analysis", International Journal of Informetrics, Vol. 4, No. 4, pp. 564-580, 2010.

[4] G. Farjamnia, Y. Gasimov, C. Kazimov, "Review of the Techniques Against the Wormhole Attacks on Wireless Sensor Networks", International Journal of Wireless Personal Communications, Vol. 105, pp. 1561-1584, 2019.

[5] S. Ji, T. Chen, S. Zhong, "Wormhole Attack Detection Algorithms in Wireless Network Coding Systems", International Journal of IEEE Transactions on Mobile Computing, Vol. 14, No. 3, pp. 660-674, 2014.

[6] E. Rodrigo, "The Physics of Stargates: Parallel Universes, Time Travel, and the Enigma of Wormhole Physics", Eridanus Press, SBN-13, 2010.

[7] A. Smith, "Black Holes: A Great Mystery", International Conference and Exposition on Engineering, Construction, Operations, and Business in Space, 2000.

[8] C.C. Aggarwal, "Data Mining", The Text Book, Switzerland, Springer, 2015.

[9] Y. Tian, D. Xu, "A Comprehensive Survey of Clustering Algorithms", International Journal of Annals of Data Science, Vol. 2, pp. 165-193, 2015.

[10] M. Vaidya, N. Dahatkar, B. Kolhe, "Predictive Monitoring System Using K-NN, QDS Classifier of Physiological Data", Information and Communication Technology for Intelligent Systems, Smart Innovation, Systems and Technologies, 2019.

[11] H. Rajaguru, S. Chakravarthy, "Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer", Asian Pacific Journal of Cancer Prevention, Vol. 20, No. 12, pp. 3777-3781, 2019.

[12] M. Connor, P. Kumar, "Fast Construction of k-Nearest Neighbor Graphs for Point Clouds", International Journal of IEEE Transactions on Visualization and Computer Graphics, Vol. 16, No. 4, pp. 599-608, 2010.

[13] S.R. Ramaswamy, R. Rastogi, K. Shim, "Efficient Algorithms for Mining Outliers from Large Data Sets", ACM SIGMOD International Conference on Management of Data, 2000.

[14] S. Guha, R. Rastogi, K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases", International Journal of ACM SIGMOD Record, Vol. 27, No. 2, pp. 73-84, 1998.

[15] S. Shen, X. Hao, Z. Liang, Y. Liu, W. Wang, L. Shao, "Real-Time Superpixel Segmentation by DBSCAN Clustering Algorithm", International Journal of IEEE Transactions on Image Processing, Vol. 25, No. 12, pp. 5933-5942, 2016.

[16] A. Barati, S.J. Dastgheib, A. Movaghar, I. Attarzadeh, "An Effective Fuzzy Based Algorithm to Detect Faulty Reading in Long Thin Wireless Sensor Networks", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 10, Vol. 4, No. 1, pp. 52-58, March 2012.

[17] L. Wang, B. Yang, Y. Chen, X. Zhang, J. Orchard, "Improving Neural-Network Classifiers Using Nearest Neighbor Partitioning", International Journal of IEEE Transactions on Neural Networks and Learning Systems, Vol. 28, No. 10, pp. 2255-2267, 2016.

[18] F. Pernkopf, M. Wohlmayr, S. Tschiatschek, "Maximum Margin Bayesian Network Classifiers", International Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, No. 3, pp. 521-532, 2011.

[19] M.J. Carey, S. Ceri, "Data-Centric System and Applications", The Text Book, Springer, 1998.

[20] F. Golabi, M. Shamsi, M.H. Sedaaghi, A. Barzegar, M. Hejazi, "Development of a New Oligonucleotide Block Location-Based Feature Extraction (BLBFE) Method for the Classification of Riboswitches", International Journal of Molecular Genetics and Genomics, Vol. 295, No. 04, pp. 525-534, 2020.

[21] P. Lopes, B. Roy, "Dynamic Recommendation System Using Web Usage Mining for E-commerce Users", International Conference on Advanced Computing Technologies and Applications, 2015.

[22] L.C. Briand, J. Wust, J.W. Daly, D.V. Poter, "Exploring the Relationships between Design Measures and Software Quality in Object-Oriented Systems", International Journal of Systems and Software, Vol. 51, No. 3, pp. 245-273, 2000.

BIOGRAPHIES

Bita Ture Savadkoohi was born in Tabriz, Iran. She obtained diploma in Software Engineering from Islamic Azad University, Iran in 2003 and the Ph.D. degree in "computer science" from University of Trento, Italy in 2010. Since 2012, she is an Assistant Professor at the Seraj Higher Education Institute, Tabriz, Iran. Her research interests included computer graphics, computational geometry (e.g. shape comparison), analysis of 3D data, software engineering, data base and data mining.

Parviz Nik Mohammadi was born in Tabriz, Iran, He received B.Sc. and M.Sc. degrees in Software Engineering from Seraj Higher Education Institute, Tabriz, Iran in 2014 and 2019, respectively. His research interests include artificial intelligence, data base system, machine learning, software engineering.