www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 12 December 2014, Page No. 9391-9394

Link Analysis in Relational Databases using Data Mining Techniques

Smita Shinde[1], Amol Rajmane[2]

[1]Department of Computer Science and Engineering
PVP Institute of Technology, Budhgaon, India
smitashinde16@gmail.com

[2]Department of Computer Science and Engineering
Ashokrao Mane Group of Institution, Vathar, *India*
amolbrajmane@gmail.com

Abstract— In data mining approach it simply assumes a random sample of different items of a single relational database. Many data mining techniques can be used to extract information from the data. Here we have proposed the work which introduces link analysis procedure discovers relationships between relational databases or graph. This work can be useful on single relational databases as well as multiple relational databases. This approach interested to analyze two or more relational databases. By taking a random walk on the database, different states can be defined. In first part relational database is displayed in terms of graph and in second part markov chain, which contains only the interested elements and preserves the characteristics of original chain and all those elements are analyzed by projecting diffusion map. Several datasets are analyzed by using the proposed methodology showing the benefits of this technique for finding relationships of relational databases or graphs.

Keywords: -Diffusion map, link analysis, markov chain, multiple relational databases, Co- sine similarity algorithm.

1. Introduction

A Real word data coming from many fields such as education, banking, marketing, social networks. . This entire research field tries to indentify links between datasets. This link analysis technique proposed to find relationships between different databases. This technique can be applied on one or more relational databases. Thousands of datasets can be analyzed by using this technique to find out different relation in the databases. By taking a random walk, different
states are studied in the databases. A link extraction technique can used to find out hidden information from the huge relational databases. This technique can be used to predict future action by studying historical data [1].

2. Link Analysis

This work is based on a markov chain which takes random walks on the different databases. It is two step actions. In first step markov chains are defined, Markov chain analyzes too many states and in second step diffusion map can be calculated. Here different datasets are analyzed by using two step procedures. All matching records are to be displayed according to top ten data mining algorithms. Here, we have used co-sine similarity algorithm. This algorithm calculates only required similar elements and in second step all the records that can be loaded into the markov chain and it shows how it is linked with other records in different databases. These two step actions extend the technique correspondence analysis. Correspondence analysis focuses on three main assistance.1.Relational Database 2. All the datasets in relational database are analyzed through markov chain 3.Reducedmarkov chain can be represented in low dimensional space through diffusion maps.

3. Need of work

- This work is intended to show the relationships between datasets of relational database.
- This work can be used in fuzzy SQL queries.

- This technique makes analysis between small instances of relational database.
- This technique provides effectiveness and scalability.
- It is away to check similarity datasets.

4. Proposed System Architecture

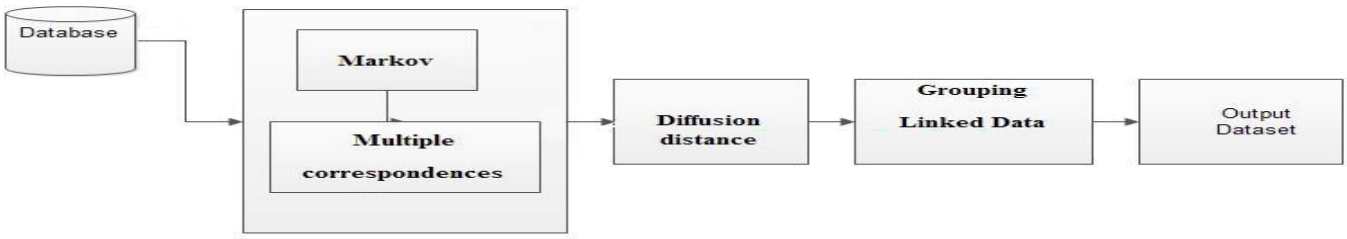

Fig A: System Architecture

5. Proposed Work

This link-analysis technique based on relational databases proposes different modules. Modules are given below

A. Registration,

B. Notation and Definition,

C. Diffusion Map,

D. Analyzing Relations,

E. Analyzing the reduced markov chain with diffusion map.

A. Registration

In this module members can be registered their personal, educational, family, company details. Here only those people can get a member id those are registered.

B. Notation and Definition

In this work, we have selected six databases .These database contains personal information, school information, college information as well as marital status and company details.

All these databases can be represented by a graph. This graph G can be defined in which elements of database are the nodes and each datasets correspondence to link for allowing to build graph from relational database

C. Diffusion Map

The Diffusion Map is a machine learning algorithm. Diffusion maps which focuses on discovering various data has been sampled from various database depending upon the local similarities. Diffusion map can be implemented by using Graph Layout Execution Engine (GLEE).GLEE is a .NET tool for laying out and visualizing directed graphs. It shows the automatic graph layout to represent complex directed graphs. It shows the number of links to the five relational databases. It can be shown in following fig B.

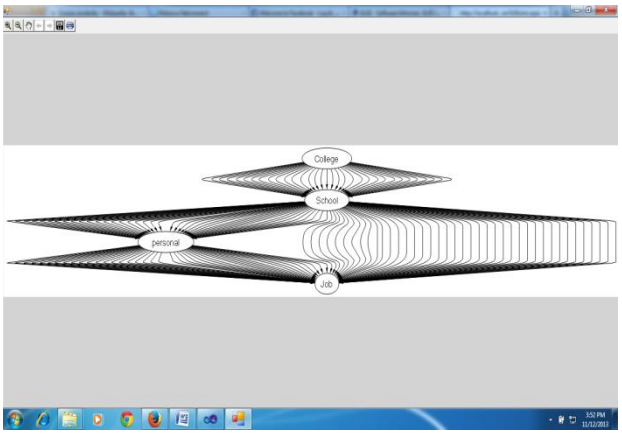

Fig B: Diffusion Map

D. Analyzing Relations

In this module the concept of stochastic complementation is studied and applied on a graph. From the initial graph, a reduced graph containing only the nodes of interest, and which is much easy to analyze and to build a graph.

E. Analyzing the reduced markov chain with diffusion map.

Markov chain is a process that consists of finite number of states and some known probabilities. This mapping allows studying the closeness between the different members. In Graph embedding it checks one of the attributes from the different relational database, such as id from registration, name of company from company details, year from college database

and gender from personal database. It gives the percentage. The resulting graph shows the proportions of the test records along with the relational database [11].

6. Algorithms Used:-

Co-sine Similarity Algorithm

Co-sine similarity is a measure of similarity between two vectors of an inner product space that measures the co-sine angle between them [8]. The co-sine of 0° is 1, and it is less than 1 for any other angle. It is a judgment for direction and not for magnitude. Two vectors with the same orientation have a Co-sine similarity of 1 and two vectors at 90° have a similarity 0 and two vectors diametrically opposed have a similarity -1 independent of their amount. Co-sine similarity is particularly used in positive space, where the outcome is neatly bounded in [0, 1]. The results are obtained by using this algorithm that is shown in fig C.

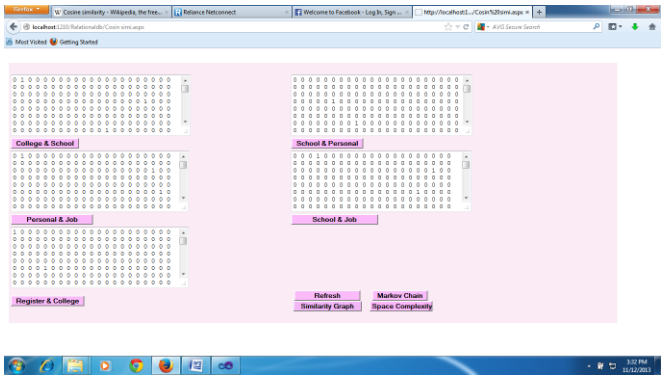

Fig C: Results of Co-sine similarity

By using similarity algorithms it shows how the numbers of links are present in the relational databases and it can be shown in the diffusion map.

7. Results

This works shows the similarity algorithm is an efficient algorithm that improves the performance between protection of sensitive information and link discovery. This algorithm is useful for analyzing large transactional databases based on a heterogeneous, multirealtional datasets described by relational databases. There is no possible way to reproduce the original database from the sanitized one. From first observation, we can say that mining more than two relational database procedure so difficult and complex to write sql query. To overcome this problem we can introduce link analysis extension for mining relational databases, here we used more than four relational databases using sql server as backend tool and Microsoft visual studio as front end tool. Microsoft visual studio is computer-based techniques used to recognize, finding, and examined business data. Our performance study compares the effectiveness and scalability of the algorithm and analyzes the fraction of connection rules to be preserved after analyzing a database. Test cases are shown in table 1.

Datasets	Old System	Our System
Datasets 1	59	8
Datasets 2	65	89
Datasets 3	49	95

Here it creates link between multiple relational databases. Once link is established between multiple relational databases then mining relational databases is very easy and efficient. The following graph shows comparison between old and new system.

Fig D: Comparison Graph

8. Conclusion and Future work

This work introduced mainly from data mining domain. The link analysis technique can be applied on large number of databases. It shows the relationships between multiple databases. This technique can be used for extracting relational databases or graph. This technique makes analysis between small instances of relational databases. This kind of data mining techniques discovers new relations in relational databases. It checks the similarity between the datasets.

Further work will be dedicated to the application on real multi relational databases. This technique focuses only on required nodes of graph or datasets of relational database [12]. This technique shows its effectiveness when it has connecting component, if there is disconnected graph or some missing tuples in databases then it is difficult to extract link from graph or database.

9. Acknowledgement

I would like to thanks my guide Prof. A. B. Rajmane for his valuable and constructive comments. I would also like to thanks Prof. Mrs. A. N. Mulla, Computer Science & Engineering Department, Annasaheb Dange College of Engineering & Technology, Ashta for her valuable support.

10. References

[1] Luh Yen , MarcoSaerens, Member, IEEE and Franois Fous "*A Link Analysis Extension of Correspondence Analysis for Mining Relational Databases*"

[2] M. Belkin and P. Niyogi, *"Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering"* Advances in Neural Information Processing Systems, vol. 14, pp. 585-591, MIT Press, 2001.

[3] M. Belkin and P. Niyogi,"*Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*" Neural Computation, vol. 15, pp. 1373-1396, 2003.

[4] F.R. Chung, "Spectral Graph Theory". Am. Math. Soc, 1997.

[5] J. Blasius, M. Greenacre, P. Groenen, and M. van de Velden, " *Special Issue on Correspondence Analysis and Related Methods*" Computational Statistics and Data Analysis, vol. 53, no. 8,pp. 3103-3106, 2009.

[6] I. Borg and P. Groenen, "*Modern Multidimensional Scaling: Theory and Applications*". Springer, 1997.

[7] P. Baldi, P. Frasconi, and P. Smyth, "*Modeling the Internet and the Web: Probabilistic Methods and Algorithms"*. John Wiley, 2003 & Sons.

[8] C. Blake, E. Keogh, and C. Merz, *"UCI Repository of Machine Learning Databases,*" Univ. California, Dept. of Information and Science, http://www.ics.uci.edu/~mlearn/.html, 1998.

[9] P. Bremaud, "*Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues"*. Springer-Verlag, 1999.

[10] F.R. Chung, *"Spectral Graph Theory",*Am. Math. Soc., 1997.

11] D.J. Cook and L.B. Holder, *"Mining Graph Data"*. Wiley and Sons, 2006.

[12] T. Cox and M. Cox, *"Multidimensional Scaling",* second ed. Chapman and Hall, 2001.

[13] N. Cressie, "*Statistics for Spatial Data".* Wiley, 1991.