A Link Analysis Technique for Mining Relational Database

[1]B.Jyothsna and [2]K.Lakshmaiah

[1] M.Tech, Dept of Computer Science and Engineering,
Madanapalle Institute of Technology and Science,Madanapalle,Chittoor (Dist),A.P,India

[2] Assoc. Professor, Dept of Computer Science and Engineering,
Madanapalle Institute of Technology and Science,Madanapalle,Chittoor (Dist),A.P,India

Abstract

This paper explains a link analysis technique allowing discovering relationships existing between elements of a relational database . More specifically, this paper is based on a random walk through the database having as many states as elements in the database. Suppose, for instance, we are interested in analyzing the relationships between elements contained in two different tables of a relational database.The two step procedure is developed for mining relational database. First, a much smaller, reduced, Markov chain, only containing the elements of interest , typically the elements contained in the two tables and preserving the main characteristics of the initial chain, is extracted by stochastic complementation. Second ,the reduced chain is analysed by projecting the states in the subspace spanned by the right Eigenvectors of the transition matrix or computing a kernel principal component analysis on a diffusion map kernel computed from the reduced graph and visualize the results by using simple or multiple correspondence analysis. The simple correspondence analysis reduces the graph when two tables are defined, and multiple correspondence analysis takes the database when there is star-schema.

Key Terms: *link analysis, diffusion map, correspondence analysis, dimensionality reduction, principal component analysis, star schema.*

1.Introduction.

Traditional statistical, machine-learning, pattern recognition, and data mining approaches usually assume a random sample of independent objects from a single relation. Many of these techniques have gone through the extraction of knowledge from data (typically extracted from relational databases), almost always leading, in the end, to the classical double entry tabular format, containing features for a sample of the population.

These features are therefore used in order to learn from the sample, provided that it is representative of the population as a whole. However, real world data coming from many fields (such as World Wide Web, marketing, social networks, or biology) are often multi-relational and interrelated. The work recently performed in statistical relational learning, aiming at working with

such datasets, incorporates research topics such as link analysis web mining social-network analysis, or graph mining.

All these research fields intend to find and exploit links between objects (in addition to features as is also the case in the field of spatial statistics, which could be of various types and involved in different kinds of relationships. The focus of the techniques has moved over from the analysis of the features describing each instance belonging to the population of interest (attribute value analysis) to the analysis of the links existing between these instances (relational analysis), in addition to the features.

This paper precisely proposes a link-analysis based technique allowing discovering relationships existing between elements of a relational database . More specifically, this work is based on a random walk through the database defining a Markov chain having as many states as elements in the database.

Suppose, for instance, we are interested in analyzing the relationships between elements contained in two different tables of a relational database. To this end, a two step procedure is developed.

First, a much smaller, reduced, Markov chain, only containing the elements of interest typically the elements contained in the two tables and preserving the main characteristics of the initial chain, is extracted by stochastic complementation. An efficient algorithm for extracting the reduced Markov chain from the large, sparse, Markov chain representing the database is proposed. Then, the reduced chain is analyzed by, for instance, projecting the states in the subspace spanned by the right eigenvectors of the transition matrix called the basic diffusion map in this paper, or by computing a kernel principal-component analysis, on a diffusion-map kernel computed from the reduced graph and visualizing the results. Indeed, a valid graph kernel based on the diffusion map distance, extending the basic diffusion map to directed graphs, is introduced.

The motivations for developing this two step procedure are two fold. First, the computation would be cumbersome, if not impossible, when dealing with the complete database. Second, in many situations, the analyst is not interested in studying all the relationships between all elements of the database, but only a subset of them.

Moreover, if the whole set of elements in the database is analyzed, the resulting mapping would be averaged out by the numerous relationships and elements we are not interested in for instance, the principal axis would be completely different. It would therefore not only reflect the relationships between the elements of interest only. Therefore, 2 reducing the Markov chain by stochastic complementation allows to focus the analysis on the elements and relationships we are

interested in. Interestingly enough, when dealing with a bipartite graph (i.e., the database only contains two tables linked by one relation), stochastic complementation followed by a basic diffusion map is exactly equivalent to simple correspondence analysis.

On the other hand, when dealing with a star schema database (one central table linked to several tables by different relations), this two-step procedure reduces to multiple correspondence analysis. The proposed methodology therefore extends correspondence analysis to the analysis of a relational database.

2.ComputingaReducedMarkovChain byStochasticComplementation.

Suppose we are interested in analyzing the relationship between two sets of nodes of interest. A reduced Markov chain can be computed from the original chain, in the following manner: First, the set of states is partitioned into two subsets, S_1—corresponding to the nodes of interest to be analyzed and S_2 corresponding to the remaining nodes, to be hidden. We further denote by n_1 and n_2 (with $n_1 + n_2 = n$) the number of states in S_1 and S_2, respectively; usually $n_2 >> n_1$. Thus, the transition matrix is repartitioned as

$$\mathbf{P}= \begin{array}{c} \\ \mathbf{s_1} \\ \mathbf{s_2} \end{array} \begin{array}{c} \begin{array}{cc} \mathbf{s_1} & \mathbf{s_2} \end{array} \\ \begin{pmatrix} \mathbf{p_{11}} & \mathbf{p_{12}} \\ \mathbf{p_{21}} & \mathbf{p_{22}} \end{pmatrix} \end{array} \longrightarrow \boxed{1}$$

The idea is to censor the useless elements by masking them during the random walk. That is, during any random walk on the original chain, only the states belonging to S_1 are recorded; all the other reached states belonging to subset S_2 being censored, and therefore, not recorded. One can show that the resulting reduced Markov chain obtained by censoring the states S_2 is the stochastic complement of the original chain [4]. Thus, performing a stochastic complementation allows to focus the analysis on the tables and elements representing the factors/features of interest.

The reduced chain inherits all the characteristics from the original chain; it simply censors the useless states. The stochastic complement Pc of the chain,partitioned as in (1),is defined as (see, for instance, [4])

$$\mathbf{P_c=p_{11}+p_{12}(I-p_{22})^{-1}p_{21}}.$$

It can be shown that the matrix Pc is stochastic, that is, the sum of the elements of each row is equal to 1 [4]; it therefore corresponds to a valid transition matrix between states of interest. We will assume that this resulting stochastic matrix is aperiodic and irreducible, that is, primitive . Indeed, Meyer showed in [4] that if the

initial chain is irreducible or aperiodic, so is the reduced chain. Moreover, even if the initial chain is periodic, the reduced chain frequently becomes aperiodic by stochastic

complementation [4]. One way to ensure the aperiodicity of the reduced chain is to introduce a small positive quantity on the diagonal of the adjacency matrix A, which
does not fundamentally change the model. Then, P has nonzero diagonal entries and the stochastic complement, Pc, is primitive (see [4], Theorem 5.1).

Let us show that the reduced chain also represents a random walk on a reduced graph Gc containing only the nodes of interest. We therefore partition the matrices
A,D,L, as

$$A = \begin{bmatrix} A11 & A12 \\ A21 & A22 \end{bmatrix}; D = \begin{bmatrix} D1 & 0 \\ 0 & D2 \end{bmatrix}; \quad L = \begin{bmatrix} L11 & L1 \\ L21 & L22 \end{bmatrix}$$

from which we easily find $Pc = D_1^{-1}(A_{11} + A_{12}(D2 - A_{22})^{-1}A_{21}) = D_1^{-1} A_c$, where we defined $A_c = (A_{11} + A12(D_2 - A_{22})^{-1}A_{21})$. Notice that if A is symmetric (the graph Gc is undirected), Ac is symmetric as well. Since Pc is stochastic, we deduce that the diagonal matrix D_1 contains the row sums of Ac and that the entries of Ac are positive. The reduced chain thus corresponds to a random walk on the graph Gc whose adjacency matrix is Ac.

Moreover, the corresponding Laplacian matrix of the graph Gc can be obtained by

$$\mathbf{L_c = D_1 - A_c = (D_1 - A_{11}) - A_{12}(D_2 - A_{22})^{-1}A_{21} = L_{11} - L_{12}L_{22}^{-1}L_{21}.}$$

since $L_{12} = -A_{12}$ and $L_{21} = -A_{21}$. If the adjacency matrix A is symmetric, L_{11} (L_{22}) is positive definite, since it is obtained from the positive semidefinite matrix L by
deleting the rows associated to S_2 (S_1) and the corresponding columns, thereby eliminating the linear relationship. Notice that Lc is simply the Schur complement of L_{22} . Thus, for an undirected graph G, instead of directly computing P_c, it is more interesting to compute L_c, which is symmetric positive definite, from which P_c
can easily be deduced: $P_c = I - D_1^{-1} L_c$, directly following from $L_c = D1 - A_c$; for a proposition of iterative computation of L_c.

3.The Diffusion Map Distance And Its Natural Kernel Matrix

Let us consider that we are given a weighted, directed, graph G possibly defined from a relational database in the following, obvious, way: each element of the database is a node and each relation corresponds to a link (for a detailed procedure allowing to build a graph from a relational database). The associated adjacency matrix A is defined in a standard way as $a_{ij} = [A]_{ij} = w_{ij}$ if node i is connected to node j and aij = 0 otherwise (say G has n nodes in total). The weight $w_{ij} > 0$ of the edge connecting node i and node j is set to have larger value if the affinity between i and j is important. If no information about the strength of relationship is available, we simply set $w_{ij} = 1$

(unweighted graph). We further assume that there are no self-loops ($w_{ii} = 0$ for i =1,….., n) and that the graph has a single connected component; that is, any node can be reached from any other node. If the graph is not connected, there is no relationship at all between the different components and the analysis has to be performed separately on each of them.

Since the Markov chain represents a random walk on the graph G, the transition matrix is simply $P = D^{-1}A$. Moreover, if the adjacency matrix A is symmetric, the Markov chain is reversible and the steady-state vector, π, is simply proportional to the degree of each state , d (which has to be normalized in order to obtain a valid probability distribution). Moreover, this implies that all the eigenvalues (both left and right) of the transition matrix are real.

4.A Link Analysis technique for mining relational database.

A link analysis procedure[6] for discovering relationships in a relational database , generalizing both simple and multiple correspondence analysis. It is based on a random walk model through the database defining a Markov chain having as many states as elements in the database. Suppose we are interested in analyzing the relationships between some elements (or records) contained in two different tables of the relational database. To this end, in a first step, a reduced, much smaller, Markov chain containing only the elements of interest and preserving the main characteristics of the initial chain is extracted by stochastic complementation. This reduced chain is then analyzed by projecting jointly the elements of interest in the diffusion-map subspace and visualizing the results. This two step procedure (see [10]) reduces to simple correspondence analysis when only two tables are defined and to multiple correspondence analyses when the database takes the form of a simple star schema.

This System proposes a link analysis based technique allowing discovering relationships existing between elements of a relational database .More specifically, this work is based on a random walk through the database having as many states as elements in the database. Suppose, for instance, we are interested in analyzing the relationships between elements contained in two different tables of a relational database. To this end, a two-step procedure is developed. First, a much smaller, reduced, Markov chain, only containing the elements of interest typically the elements contained in the two tables and preserving the main characteristics of the initial chain, is extracted by stochastic complementation.

In this paper we have two main contributions:

1. A two step procedure for analyzing weighted graphs or relational databases is proposed.
2. It is shown that the suggested procedure extends correspondence analysis.

5.Analysing the reduced markov chain links with correspondence analysis.

Once a reduced Markov chain containing only the nodes of interest has been obtained,one may want to visualize the graph in a low dimensional space preserving as accurately as possible the proximity between the nodes. This is the second step of our procedure. For this purpose, we propose to use the diffusion maps. Interestingly enough, computing a basic diffusion map on the reduced Markov chain is equivalent to correspondence analysis in two special cases of interest: a bipartite graph and a star-schema database. Therefore, the proposed two step procedure can be considered as a generalization of correspondence analysis.

Correspondence analysis (see,for instance,[5] is a widely used multivariate statistical analysis technique which still is the subject of much research efforts simple correspondence analysis aims to provide insights into the dependence of two categorical variables. The relationships between the attributes of the two categorical variables are usually analyzed through a biplot [5] a 2D representation of the attributes of both variables. The coordinates of the attributes on the biplot are obtained by computing the eigenvectors of a matrix. Many different derivations of simple correspondence analysis have been developed, allowing for different interpretations of the technique, such as maximizing the correlation between two discrete variables, reciprocal averaging, categorical discriminant analysis, scaling and quantification of categorical variables, performing a principal component analysis based on the chi-square distance, optimal scaling, dual scaling, etc.Multiple correspondence analysis is the extension of simple correspondence analysis to a larger number of categorical variables.

6. Simple Correspondence Analysis

Simple correspondence analysis aims to study the relationships between two random variables x_1 and x_2 (the features) having each mutually exclusive, categorical,outcomes, denoted as attributes. Suppose the variable x_1 has n_1 observed attributes and the variable x_2 has n_2 observed attributes, each attribute being a possible outcome value for the feature. An experimenter makes a series of measurements of the features x_1; x_2 on a sample of v_g individuals and records the outcomes in a frequency (also called contingency) table, f_{ij}, containing the number of individuals having both attribute $x_1 = i$ and attribute $x_2 = j$. In our relational database, this corresponds to two tables, each table corresponding to one variable, and containing the set of observed attributes (outcomes) of the variable. The two tables are linked by a single relation (see Fig. 1 for a simple example). This situation can be modeled as a bipartite graph, where each node corresponds to an attribute and links are only defined between attributes of x_1 and attributes of x_2. The weight associated to each link is set to $w_{ij} = f_{ij}$, quantifying the strength of the relationship between i and j.

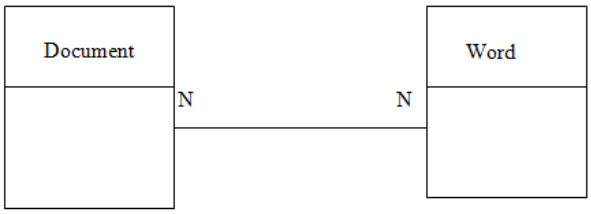

Fig. 1. Trivial example of a single relation between two variables,Document and Word. The Document table contains outcomes of documents while the Word table contains outcomes of words.

The associated n X n adjacency matrix and the corresponding transition matrix can be factorized as

$$A=\begin{bmatrix} O & A12 \\ A21 & O \end{bmatrix} \quad , P = \begin{bmatrix} 0 & P12 \\ P21 & 0 \end{bmatrix},$$

where O is a matrix full of zeroes.

Suppose we are interested in studying the relationships between the attributes of the first variable x_1, which corresponds to the n_1 first elements. By stochastic complementation (see (10)), we easily obtain $Pc = P_{12}P_{21} = D_1^{-1}A_{12}D_2^{-1}A_{21}$.

Computing the diffusion map for t = 1aims to extract the subdominant right-hand eigenvectors of Pc, which exactly corresponds to correspondence analysis (see, for instance, [2], (4.3.5)). Moreover, it can easily be shown that Pc has only real nonnegative eigen values, and thus, ordering the eigen values by modulus is equivalent to ordering them by value. In correspondence analysis, eigen values reflect the relative importance of the dimensions: each eigen value is the amount of inertia a given dimension exhibits in the frequency table . The basic diffusion map after stochastic complementation on this bipartite graph therefore leads to the same results as simple correspondence analysis. Relationships between simple correspondence analysis and link analysis techniques have already been highlighted. For instance,

Zha et al. [6] showed the equivalence of a normalized cut performed on a bipartite graph and simple correspondence analysis. On the other hand, Saerens et al. investigated the relationships between Kleinberg's HITS algorithm, and correspondence analysis or principal component analysis [7].

7. Multiple Correspondence Analysis

Multiple correspondence analysis assigns a numerical score to each attribute of a set of p > 2 categorical variables[4]. Suppose the data are available in the form of a star-schema: the individuals are contained in a main table and the categorial features of these individuals, such as education level, gender, etc., are contained in p auxiliary, satellite, tables. The corresponding graph is built

naturally by defining one node for each individual and for each attribute while a link between an individual and an attribute is defined when the individual possesses this attribute. This configuration is known as a star-schema in the data warehouse or relational database fields (see Fig. 2 for a trivial example).

Let us first renumber the nodes in such a way that the attribute nodes appear first and the individuals nodes last. Thus, the attributes-to-individuals matrix will be denoted by A_{12}; it contains a 1 on the (i,j) entry when the individual j has attribute i, and 0 otherwise. The individuals-to-attributes matrix, the transpose of the attributes-to-individuals matrix, is A_{21}. Thus, the adjacency matrix of the graph is

$$\mathbf{A}=\begin{bmatrix} \mathbf{O} & \mathbf{A12} \\ \mathbf{A21} & \mathbf{O} \end{bmatrix} .$$

Now, the individuals-to-attributes matrix exactly corresponds to the data matrix $A_{21} = X$ containing, as rows, the individuals and, as columns, the attributes. Since the different features are coded as indicator (dummy) variables.

Assuming binary weights, the matrix D_1 contains on its diagonal the frequencies of each attribute, that is, the number of individuals having this attribute. On the other hand, D_2 contains p on each element of its diagonal, since each individual has exactly one attribute for each of the p features(attributes corresponding to a feature are mutually exclusive). Thus, $D_2 = P.I$ and $P_{12} = D_1^{-1} A_{12}$, $P_{21} = D_2^{-1} A_{21}$.

Each table contains outcomes of the corresponding random variable. A row of the X matrix contains a 1 if the individual has the

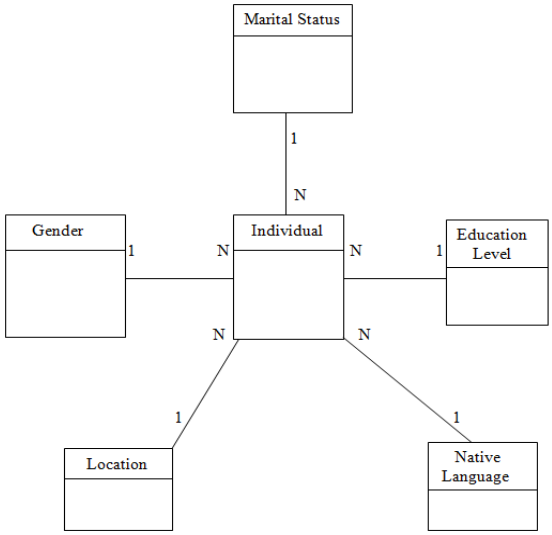

Fig. 2. Trivial example of a star-schema relation between a main variable, Individual, and auxiliary variables, Gender, Education level, etc.

corresponding attribute and 0 otherwise. We thus have $A_{21} = X$ and $A_{12} = X^T$.

Suppose we are first interested in the relationships between attribute nodes, thereby hiding the individual nodes contained in the main table. By stochastic complementation , the corresponding attribute-attribute transition matrix is

$$\mathbf{P_c=D_1^{-1}A_{12}D_2^{-1}A_{21}=1/p.D_1^{-1}A_{12}A_{21}=1/p.D_1^{-1}X^TX=1/p.D_1\text{-}^1F,}$$

where the element f_{ij} of the frequency matrix $F= X^TX$, also called the Burt matrix, contains the number of cooccurences of the two attributes i and j, that is, the number of individuals having both attribute i and attribute j.

The largest nontrivial right eigenvector of the matrix P_c represents the scores of the attributes in a multiple correspondence analysis. Thus, computing the eigen values and eigenvectors of P_c and displaying the nodes with coordinates proportional to the eigenvectors, weighted by the corresponding eigen value, exactly corresponds to multiple correspondence analysis. This is precisely what we obtain when computing the basic diffusion map on P_c with t =1. Indeed, as for simple correspondence analysis, it can easily be shown that Pc has real nonnegative eigen values, and thus, ordering the eigen values by modulus is equivalent to ordering by value.

If we are interested in the relationships between elements of the main table (the individuals)instead of the attributes, we obtain

$$\mathbf{P_c=1/p.A_{21}D_1^{-1}A_{12}=1/p.XD_1^{-1}X^T,}$$

Which , once again, is exactly the result obtained by multiple correspondence analysis.

8.Experiments.

Mining relational database were performed for entering data for personnel details like personal,educational,family,business details.In this paper to maintain population. So, each & every person fulfills the details.

9.Conclusion

This paper introduced a link analysis based technique allowing to analyze relationships existing in relational database. The database is viewed as a graph where the nodes correspond to the elements contained in the tables and the links correspond to the relations between the tables. A two step procedure is defined for analyzing the relationships between elements of interest contained in a table, or a subset of tables. More precisely, this work (1) proposes to use stochastic complementation for extracting a subgraph containing the elements of interest from the original graph and (2) introduces a kernel based extension of the basic diffusion map for displaying and analyzing the reduced subgraph. It is shown that the resulting method is closely related to correspondence analysis. Several datasets are analyzed by using this procedure, showing that it

seems to be well-suited for analyzing relationships between elements. Indeed, stochastic complementation considerably reduces the original graph and allows to focus the analysis on the elements of interest, without having to define a state of the Markov chain for each element of the relational database. However, one fundamental limitation of this method is that the relational database could contain too many disconnected components, in which case our linkanalysis approach is almost useless. Moreover, it is clearly not always an easy task to extract a graph from a relational database, especially when the database is huge. These are the two main drawbacks of the proposed two-step procedure. Further work will be devoted to the application of this methodology to fuzzy SQL queries or fuzzy information retrieval. The objective is to retrieve not only the elements strictly complying with the constraints of the SQL query, but also the elements that almost comply with these constraints and are therefore close to the target elements. We will also evaluate the proposed methodology on real relational databases.

References:-

1. Floris Geerts and Heikki Mannila and Evimaria Terzi, "Relational link-based ranking" , Laboratory for Foundations Basic Research Unit of Computer Science Helsinki Institute for Information Technology
2. M.J.Greenacre, "Theory and Applications of Correspondence Analysis".Academic Press, 1984.
3. Graham Williams, Markus Hegland and Stephen Roberts," A Data Mining Tutorial".
4. C.D.Meyer,"Stochastic complementation,Uncouling markovchains, and the theory of nearly reducible systems".
5. Oleg Nenadi´c and Michael Greenacre, "Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package".
6. Hongyuan Zha and Chris Ding and Horst Simon," Bipartite Graph Partitioning and Data Clustering".
7. Marco Saerens & Francois Fouss, "HITS is PCA",
8. Herv'eAbdi,"Multivariate Analysis".
9. Han and kamber,"Data Mining concepts and Techniques".
10. Heungsun Hwang and William R. Dillon ,"An Extension of Multiple Correspondence Analysis for Identifying Heterogeneous Subgroups of Respondents".