Fuzzy Information Retrieval from Mining Relational Database by Using Link Analysis Mining Methods

[1]M. Sivanjaneyulu, [2]A. Anuradha

[1,2]DVR & Dr. HS MIC College of Technology, Kanchikacherla, Krishna, AP, India

Abstract

Link Analysis algorithms have been powering various search engines for efficient web information retrievals. Instead of web, in this paper we propose to use Link Analysis as an extension of correspondence analysis in a relational database for its ability to effectively discover relationships. Initially, a reduced, much smaller, Markov chaining containing only the elements of interest is extracted and refined by stochastic complementation. This reduced chain is then analyzed by projecting jointly the elements (entity relations in relational database) of interest in a kernel version of the diffusion-map subspace along with spectral clustering to visualize the results. Also applying this technique for fuzzy information retrievals can improve overall performance in a relational database. Experiments show the usefulness of the technique for extracting relationships in relational databases.

Keywords

Fuzzy, Link Analysis, Stochastic Complementation, Diffusion-Map Subspace, Markov Chaining

I. Introduction

Traditional statistical, machine-learning, pattern recognition, and data-mining approaches usually assume a random sample of independent objects from a single relation. 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 The analysis of cross-referencing patterns—“link analysis”—has come to play an important role in modern information retrieval. Link analysis algorithms have been successfully applied to web hyperlink data to identify authoritative information sources, and to academic citation data to identify.

In this paper,we proposes a link-analysis based technique allowing to discover relationships existing between elements of a relational database. The technique is based on a random-walk through the database defining a Markov chain having as many states as elements in the database and a two-step procedure is developed. First, a much smaller, reduced, Markov chain, only containing the elements of interest 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.

In this paper, we also applying this technique for fuzzy SQL queries or fuzzy information retrieval which improve overall performance. 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.

II. Notations and Definitions

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. The associated adjacency matrix A is defined in a standard way as aij = $[A]_{ij} = w_{ij}$ if node i is connected to node j and $a_{ij} = 0$ otherwise. 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.

We define a natural random walk through the graph in the usual way by associating a state to each node and assigning a transition probability to each link. Thus, a random walker can jump from element to element and each element therefore represents a state of the Markov chain describing the sequence of visited states. A random variable s(t) contains the current state of the Markov chain at time step t: if the random walker is in state i at time t, then s(t) = i. The transition probabilities only depend on the current state and not on the past ones (first-order Markov chain). Since the graph is completely connected, the Markov chain is irreducible, that is, every state can be reached from any other state. we define P as the transition matrix with entries pij , the evolution of the Markov chain is characterized by x(t + 1) =PT x(t), with x(0) = x_0 and T is the matrix transpose.

III. Diffusion-Map Distance

In our two-step procedure, a diffusion-map projection, based on the so-called diffusion-map distance, will be performed after stochastic complementation. Since P is aperiodic, irreducible and reversible, it is wellknown that all the eigenvalues of P are real and the eigenvectors are also real. Distance between states i and j,

$$d_{ij}^2(t) = \sum_{k=1}^{n} \frac{(x_{ik}(t) - x_{jk}(t))^2}{\pi_k}$$

$$\propto (\mathbf{x}_i(t) - \mathbf{x}_j(t))^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{x}_i(t) - \mathbf{x}_j(t))$$

since, for a simple random walk on an undirected graph, the entries of the steady-state vector Π are proportional to the generalized degree of each node. This distance, called the diffusion-map distance, corresponds to the sum of the squared differences between the probability distribution of being in any state after t transitions when starting (i.e., at time t = 0) from two different states, state i and state j.

A kernel view of the diffusion-map distance extension presents several advantages in comparison with the original basic diffusion map:

1. the kernel version of the diffusion map is applicable to directed graphs while the original model is restricted to undirected graphs
2. the extended model induces a valid kernel on a graph
3. the resulting matrix has the nice property of being symmetric positive definite – the spectral decomposition can thus be computed on a symmetric positive definite matrix, and finally.
4. the resulting mapping is displayed in a Euclidean space in which the coordinate axis are set in the directions of maximal

variance by using kernel

$$d_{ij}^2(t) = (\mathbf{x}_i(t) - \mathbf{x}_j(t))^{\mathrm{T}} \mathbf{W} (\mathbf{x}_i(t) - \mathbf{x}_j(t))$$

principal- component analysis or multidimensional scaling. This kernel-based technique will be referred to as the diffusion-map kernel PCA or the KDM PCA. The diffusion-map distance is therefore redefined as referred to as the diffusion-map kernel.

IV. Reduced Markov Chain Computation

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, S1 – corresponding to the nodes of interest to be analyzed – and S2 – corresponding to the remaining nodes, to be hidden. During any random walk on the original chain, only the states belonging to S1 are recorded; all the other reached states belonging to subset S2 being censored and therefore not recorded. One can show that the resulting reduced Markov chain obtained by censoring the states S2 is the stochastic complement of the original chain. 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.

V. Analyzing the Reduced Markov Chain

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. 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 starschema database. Therefore, the proposed two-step procedure can be considered as a generalization of correspondence analysis.

Correspondence analysis 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 – a two-dimensional 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 components 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.

VI. Fuzzy Information Retrival

To retrieve the information, fuzzy matching is performed on reference tables. Fuzzy matching is used by fuzzy look up transformations. The Fuzzy Lookup transformation differs from the Lookup transformation in its use of fuzzy matching. The Lookup transformation uses an equi-join to locate matching records in the reference table. It returns either an exact match or nothing from the reference table. In contrast, the Fuzzy Lookup transformation uses fuzzy matching to return one or more close matches from the reference table. A Fuzzy Lookup transformation frequently follows a Lookup transformation in a package data flow.

First, the Lookup transformation tries to find an exact match. If it fails, the Fuzzy Lookup transformation provides close matches from the reference table. For instance, if the input string is SMITH, I want to retrieve all similar results, such as SMYTH, AMITH, SMITH, SMYTHE, etc., ideally with a measure of match closeness, e.g., 98%. In general, a fuzzy set fz (of a given variable z) is formally specified as fz = {(x, ufz (x))|x ϵ Uz, ufz (x) ϵ [0, 1]}, where ufz (x) ϵ [0, 1] is the membership function of fz. With the fuzzy approach, this variation is generally captured through degrees that belongs to distinct value sets. As a result, the given measure may be regarded as 'Moderate' or 'High', but with different membership degrees.

VII. Conclusion

A link-analysis based technique allowing to analyze relationships existing in relational databases. 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 (1) . First, a much smaller, reduced, Markov chain, only containing the elements of interest and preserving the main characteristics of the initial chain, is extracted by stochastic complementation (2) by computing a kernel principal-component analysis on a diffusion-map kernel computed from the reduced graph and visualizing the results.

Several datasets are analyzed by using this procedure, showing that it seems to be well-suited for analyzing relationships between elements. In this paper, fuzzy SQL queries or fuzzy information retrieval is used along with link analysis based techinque. The objective is to retrieve not only the elements strictly complying with the constraints of the SQL query, but also the elements that are close to the target elements. Applying both, link analysis based technique for fuzzy information retrievals can improve overall performance in a relational database.

References

[1] M. Thelwall,"Link Analysis: An Information Science Approach", Elsevier, 2004.

[2] F. Fouss, A. Pirotte, J.-M. Renders, M. Saerens,"Random-walk computation of similarities between nodes of a graph, with application to collaborative recommendation", IEEE Transactions on Knowledge and Data Engineering, 19(3), pp. 355–369, 2007.

[3] T. Boongoen, Q. Shen, C. Price,"Fuzzy qualitative link analysis for academic performance evaluation", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 19(3), pp. 559–585, 2011.

[4] I. Fellegi, A. Sunter,"Theory of record linkage", Journal of the American Statistical Association, 64, pp. 1183–1210, 1969.

[5] J. Blasius, M. Greenacre, P. Groenen, M. van de Velden, "Special journal issue on correspondence analysis and related methods", Computational Statistics and Data Analysis, 53(8), pp. 3103–3106, 2008.

[6] F. Fouss, J.-M. Renders, M. Saerens,"Links between Kleinberg's hubs and authorities, correspondence analysis and Markov chains", In Proceedings of the 3th IEEE International Conference on Data Mining (ICDM), pp. 521–524, 2003.

[7] F. Geerts, H. Mannila, E. Terzi,"Relational link-based ranking", Proceedings of the 30th Very Large Data Bases Conference (VLDB), pp. 552–563, 2004.

[8] S. Lafon, A. B. Lee,"Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization", IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(9), pp. 1393–1403, 2006.