ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

Analyzing the Reduced Markov Chain with the Basic Diffusion Map

[1]R. Sowmya, [2]J. Srinivasa Rao, [3]Dr. K. Rama Krishnaiah
[1,2,3]Dept. of CSE, NOVA College of Engineering & Tech., Jupudi, Ibrahimpatnam Vijayawada, AP, India

Abstract
This work introduces a link-analysis procedure for discovering relationships in a relational database or a graph, 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 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. On the other hand, a kernel version of the diffusion-map distance, generalizing the basic diffusion-map Distance to directed graphs, is also introduced and the links with spectral clustering are discussed. Several datasets are analyzed by using the proposed methodology, showing the usefulness of the technique for extracting relationships in relational databases or graphs.

Keywords
Graph mining, link analysis, kernel on a graph, diffusion map, correspondence analysis, dimensionality reduction, Statistical relational learning

I. Introduction

Many data mining (i.e. knowledge discovery) techniques are these days in use for ontology learning – text mining, Web mining, graph mining, network analysis, link analysis, relation data mining, stream mining, and so on [6]. In the current state-ofthe- art bundle mining of software code and the associated documentations is not explicitly addressed. With the growing amounts of software, especially open-source software libraries, we argue that mining such data is worth considering as a new methodology. Thus we introduce the term "software mining" to refer to such methodology. The term denotes the process of extracting knowledge (i.e. useful information) out of data sources that typically accompany an open-source software library. The motivation for software mining comes from the fact that the discovery of reusable software artifacts is just as important as the discovery of documents and multimedia contents. According to the recent Semantic Web trends, contents need to be semantically annotated with concepts from the domain ontology in order to be discoverable by intelligent agents. Because the legacy content repositories are relatively large, cheaper semi-automatic means for semantic annotation and domain ontology construction are preferred to the expensive manual labor. Furthermore, when dealing with software artifacts it is possible to go beyond discovery and also support other user tasks such as composition, orchestration, and execution. The need for ontology-based systems has yield several research and development projects supported by EU that deal with this issue. One of these projects is TAO (http://www.tao-project.eu) which stands for Transitioning Applications to Ontologies. In this paper we present work in the context of software mining for the domain ontology construction. We illustrate the proposed approach on the software mining case study based on GATE [3], an open-source software library for naturallanguage processing written in Java programming language. We interpret "software mining" as being a combination of methods for structure mining and for content mining. To be more specific, we approach the software mining task with the techniques used for text mining and link analysis. The GATE case study serves as a perfect example in this perspective. On concrete examples we discuss how each instance (i.e. a programming construct such as a class or a method) can be represented as a feature vector that combines the information about how the instance is interlinked with other instances, and the information about its (textual) content. The so-obtained feature vectors serve as the basis for the construction of the domain ontology with OntoGen [4], a system for semi-automatic, data-driven ontology construction, or by using traditional machine learning algorithms such as clustering, classification, regression, or active learning.

II. Existing System

We now introduce1 a variant of the basic "diffusion map" model introduced by Nadler et al. and Pons and Latapy, which is still well-defined when the original graph is directed. In other words, we do not assume that the initial adjacency matrix A is symmetric in this section. This 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 axes are set in the directions of maximal variance by using (uncentered if the kernel is not centered) kernel principal component analysis or multidimensional scaling [6,12]. This kernel-based technique will be referred to as the diffusion map kernel PCA or the KDM PCA. Let us defineW¼ ðDiagð_ÞÞ_1, where _ is the stationary distribution of the finite Markov chain. Remember that if the adjacency matrix is symmetric, the stationary distribution of the natural random walk is proportional to the degree of the nodes, W/ D_1. The diffusion map distance is therefore redefined as d2 , (6) becomes where we defined referred to as the diffusion map kernel. Thus, the matrix KDM is the natural kernel (inner product matrix) associated to the squared diffusion map distances [6], [12]. It is clear that this matrix is symmetric positive semidefinite and contains inner products in a euclidean space, where the node vectors are exactly separated by dijðtÞ (the proof is straightforward and can be found in [17]—appendix D— where the same reasoning was applied to the commute time kernel). It is therefore a valid kernel matrix. erforming a (uncentered if the kernel is not centered) principal component analysis (PCA) in this embedding space aims to change the coordinate system by putting the new coordinate axes in the directions of maximal variances. From the theory of

classical multidimensional scaling [6,12], it suffices2 to compute the m first eigenvalues/ eigenvectors of KDM and to consider that these eigenvectors multiplied by the squared root of the corresponding eigenvalues are the coordinates of the nodes in the principal component space spanned by these eigenvectors (see [6,12]; for a similar application with the commute time kernel, for the general definition of kernel PCA). In other words, we compute the m first eigenvalues/ eigenvectors of KDM : KDMwk ¼ _kwk, where the wk are orthonormal. Then, we represent each node i in a mdimensional euclidean space with coordinates . corresponds to element i of the eigenvector wk associated to eigenvalue _k this is the vector representation of state i in the principal component space. It can easily be shown that when the initial graph is undirected, this PCA based on the kernel matrix KDM is similar to the diffusion map introduced in the last section, up to an isometry. Indeed, by the classical theory of multidimensional scaling [6,12], the eigenvectors of the kernel matrix KDM multiplied by the squared root of the corresponding eigenvalues define coordinates in a euclidean space, where the observations are exactly separated by the distance dijðtÞ. Since this is exactly the property of the basic diffusion map (3), both representations are similar up to an isometry. Notice that the resulting kernel matrix can easily be centered by HKDMH with H ¼ I _ ðeeT=nÞ, where e is a column vector all of whose elements are "1" (i.e., e ¼ ½1; 1; . . . ; 1_T). H is called the centering matrix. This aims to place the origin of the coordinates of the diffusion map at the center of gravity of the node vectors.

III. Proposed Work

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. 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 twostep 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. As stated, for instance, in 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 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 thetechnique, 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 As the extension of simple correspondence analysis to a larger number of categorical variables. Zachary has studied the relations between the 34 members of a karate club. A disagreement between the club instructor (node 1) and the administrator (node 34) resulted in the split of the club into two parts. Each member of the club is represented by a node of the graph and the weight between nodes ði; jÞ is set to be the number of times member i and member j met outside the club. The built friendship network between members allows to discover how the club split, but its mapping also allows to study the proximity between the different members. It can be observed from the graph embeddings provided by the basic diffusion map that the value of parameter t has no influence on the nodes' position, but only on the axis scaling (remember (4)). On the contrary, thE influence of the value of parameter t is clearly visible on the KDM PCA mapping. However, the projection of a graph on a 2D space usually leads to a loss of information. The information preservation ratio can be estimated using ð_1 þ _2Þ= P i _i; it was computed for the 2D mapping of the network using the basic diffusion map and the KDM PCA. It can be observed that the ratio increases with t but, overall, the information is better preserved with the KDM PCA than with the basic diffusion map. This is due to a better choice of the projection axes for the KDM PCA that are oriented in the directions of maximal variance. For the KDM PCA, the choice of the value for the parameter t depends on the graph's property that the user wishes to highlight. When t is low (t ¼ 1), nodes with few relevant connections are considered as outliers and rejected from the community. On the contrary, when t increases, the effect of marginal nodes fades and only the global members structure, similar to the basic diffusion map, is visible. The unweighted graph represents the associations between bottlenose dolphins from the same community, observed over a period of 7 years. Only the basic diffusion map with t ¼ 1 is shown (since, once again, the parameter t has no influence on the mapping) while the KDM PCA is computed for t ¼ 1; 3, and 6. It can be observed from KDM PCA that a dolphin (the 61th member, "Zig") is rejected far away from the community due to his lack of interaction with other dolphins. His only contact is also poorly connected with the remaining community. As explained the basic diffusion map and the KDM PCA mappings become similar when t increases. For instance, when t ¼ 6 or more, the outlier is no more highlighted and only the main structure of the network is visible. Two main subgroups can be identified from the mapping (notice that Newman and Girvan actually identified four subcommunities by clustering). The graph is composed of four objects (e_1; e_2; e_3, and e_4) belonging to two different classes (c_1 and c_2). Each object is also connected to one or many of the five attributes (a_1; . . . ; a_5). The reduced graph mapping obtained by our two-step procedure allows to highlight the relations between the attribute values (i.e., the a nodes) and the classes (i.e., the c nodes). To achieve this goal, the e nodes are eliminated by performing a stochastic complementation: only the a and c nodes are kept. The resulting subgraph is displayed on a 2D plane by performing a KDM PCA. Since the connectivity between nodes a_1 and a_2 (a_3; a_4, and a_5) is larger than with the remaining nodes, these two (three) nodes are close together on the resulting map. Moreover, node c1 (c_2) is highly connected to nodes a_1; a_2 ða_3; a_4, and a_5Þ through indirect links and is therefore displayed close to these nodes.

IV. Results

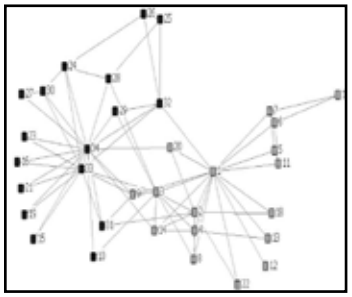


Fig. 1:

ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

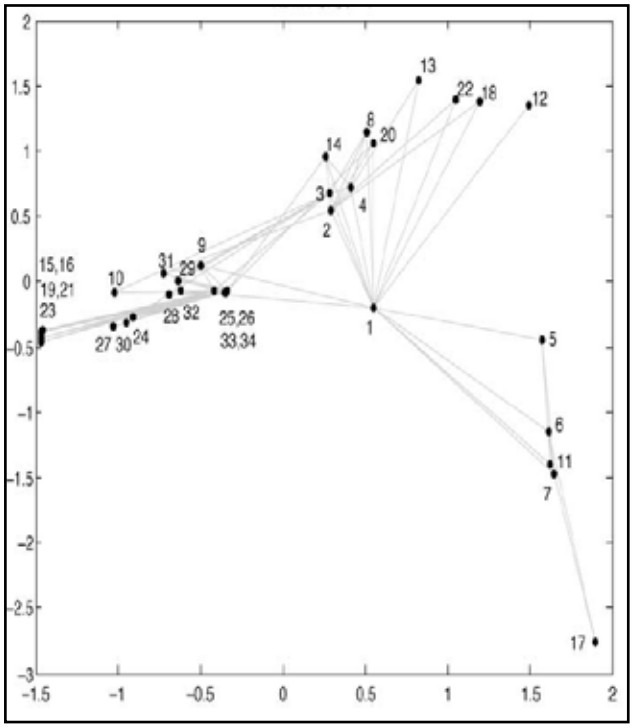


Fig. 2:

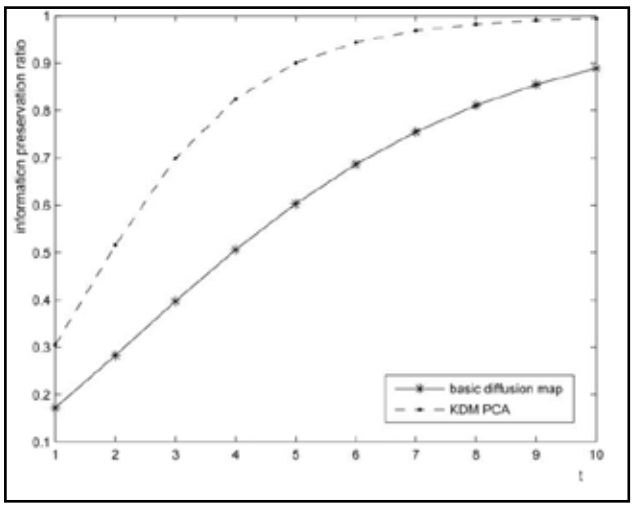


Fig. 3:

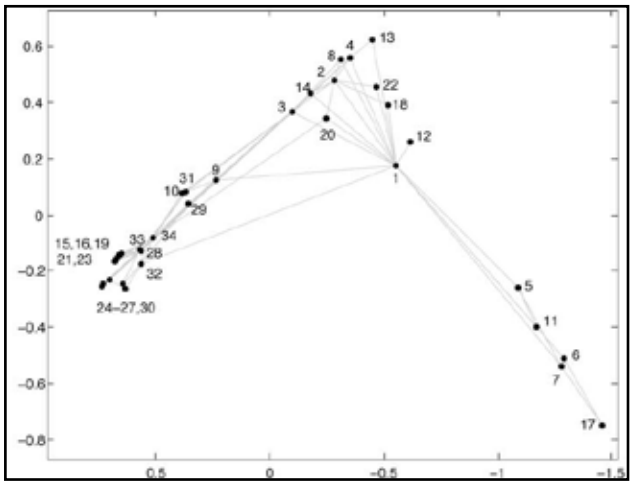


Fig. 4:

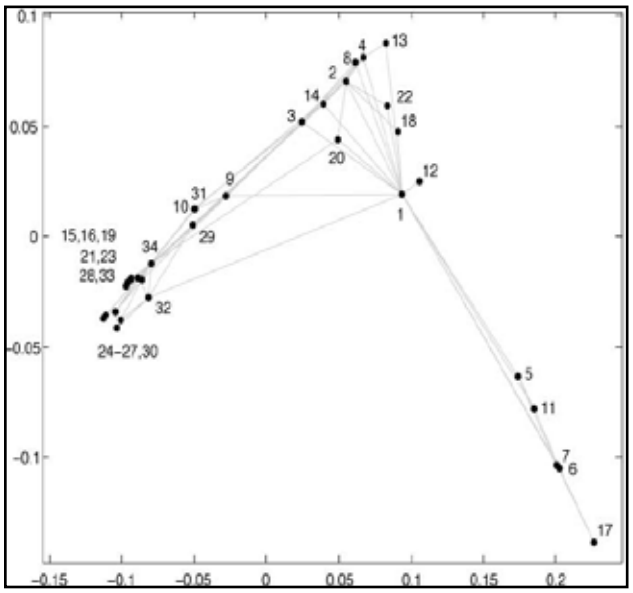


Fig. 5:

V. Conclusions

This work introduced 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. 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 ethod is closely related to correspondence analysis

References

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

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

[4] C. Blake, E. Keogh, C. Merz (1998),"UCI Repository of Machine Learning Databases", Univ. California, Dept. of Information and Computer Science, [Online] Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[5] J. Blasius, M. Greenacre, P. Groenen, 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, P. Groenen,"Modern Multidimensional Scaling: Theory and Applications", Springer, 1997.

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

[8] P. Carrington, J. Scott, S. Wasserman,"Models and Methods in Social Network Analysis", Cambridge Univ. Press, 2006.

[9] S. Chakrabarti,"Mining the Web: Discovering Knowledge from Hypertext Data", Elsevier Science, 2003.

[10] F.R. Chung,"Spectral Graph Theory", Am. Math. Soc., 1997.
[11] D.J. Cook, L.B. Holder,"Mining Graph Data", Wiley and Sons, 2006.
[12] T. Cox, M. Cox,"Multidimensional Scaling", second ed. Chapman and Hall, 2001.
[13] N. Cressie,"Statistics for Spatial Data", Wiley, 1991.
[14] P. Demartines, J. Herault,"Curvilinear Component Analysis: A Self-Organizing Neural Network for Nonlinear Mapping of Data Sets", IEEE Trans. Neural Networks, Vol. 8, No. 1, pp. 148-154, Jan. 1997.
[15] C. Ding,"Spectral Clustering", Tutorial presented at the 16th European Conf. Machine Learning (ECML' 05), 2005.
[16] P. Domingos,"Prospects and Challenges for Multi-Relational Data Mining", ACM SIGKDD Explorations Newsletter, Vol. 5, No. 1, pp. 80-83, 2003.
[17] 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 Trans. Knowledge and Data Eng., Vol. 19, No. 3, pp. 355-369, Mar. 2007.
[18] F. Fouss, J.-M. Renders, M. Saerens,"Links between Kleinberg's Hubs and Authorities, Correspondence Analysis and Markov Chains", Proc. Third IEEE Int'l Conf. Data Mining (ICDM), pp. 521-524, 2003.



R.Sowmya has completed B.Tech from Sarada Institute Of Tech and Sciences Khammam in 2010 and pursuing M.Tech in NOVA College of Engineering Technology for Women. Her research work includes concepts of Data Mining and Data Warehousing used for fast access of data stored in databases.



J.Srinivasa Rao received his M.Tech in Computer Science and Engineering from KLC University in 2008.He is a very good administrator& researcher who has presented nearly 10 various International Journals and also attended for National and International Conferences. He worked in Prestigious K L University in the Department of Computer Science and Engineering. He is a good researcher who has worked mostly on neural networks, network security.



Dr. K. Rama Krishnaiah is a highly qualified person holding M.Tech and Ph.D degree in CSE, an efficient and eminent academician. He is an outstanding administrator, a prolific researcher, who has published 21 research papers in various international journals and a forward looking educationist. He worked in prestigious K L University for 13 years and he contributed his service for NBA accreditation in May 2004, Aug 2007 with 'record rating', ISO 9001:2000 in 2004, Autonomous status in 2006, NAAC accreditation of UGC in 2008 and University status in 2009.