Efficiently Reduce the Relationships among the Database Tables using Markov Chain and Diffusion Map

[1]G. Tatayyanaidu, [2]K. T. V. Subba Rao, [3]M. Bala Krishna

[1,2,3]Dept. of CSE, Akula Gopayya College of Engineering and Technology, Tadepalligudem, AP, India

Abstract

when the database tables or nodes in a graph contain more than one relationship, it will create complexity. This paper precisely proposes a link-analysis-based technique allowing discovering relationships existing between elements of a relational database or, more generally, a graph. 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. 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 databaseAn 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

Keywords

Graph Mining, Link Analysis, Kernel on a Graph, Diffusion Map, Correspondence Analysis, Dimensionality Reduction, Statistical Relational Learning

I. Introduction

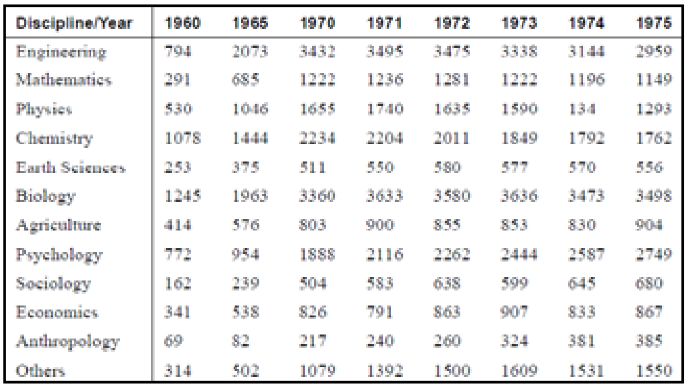
Graph Mining-Structure mining or structured data mining is the process of finding and extracting useful information from semi structured data sets. Graph mining is a special case of structured data mining. Graph mining is extracting the knowledge of data by graphs.

Simple correspondence analysis, a way of analyzing a two-way table of data. Correspondence analysis is a statistical visualization method for picturing the associations between the levels of a two-way contingency table. The name is a translation of the French Analyses des Correspondences, where the term correspondence denotes a "system of associations" between the elements of two sets.

In a two-way contingency table, the observed association of two traits is summarized by the cell frequencies, and a typical inferential aspect is the study of whether certain levels of one charactertistic are associated with some levels of another. Correspondence analysis is a geometric technique for displaying the rows and columns of a two-way contingency table as points in a low-dimensional space, such that the positions of the row and column points are consistent with their associations in the table. The goal is to have a global view of the data that is useful for interpretation.

To illustrate correspondence analysis, consider the multidimensional time series on the number of science doctorates conferred in the USA from 1960 to 1975 that is shown in Table 1 (Greenacre, 1984). Correspondence analysis of these data yields

Table 1:



Discipline/Year	1960	1965	1970	1971	1972	1973	1974	1975
Engineering	794	2073	3432	3495	3475	3338	3144	2959
Mathematics	291	685	1222	1236	1281	1222	1196	1149
Physics	530	1046	1655	1740	1635	1590	134	1293
Chemistry	1078	1444	2234	2204	2011	1849	1792	1762
Earth Sciences	253	375	511	550	580	577	570	556
Biology	1245	1963	3360	3633	3580	3636	3473	3498
Agriculture	414	576	803	900	855	853	830	904
Psychology	772	954	1888	2116	2262	2444	2587	2749
Sociology	162	239	504	583	638	599	645	680
Economics	341	538	826	791	863	907	833	867
Anthropology	69	82	217	240	260	324	381	385
Others	314	502	1079	1392	1500	1609	1531	1550

Measure of similarity between the row-frequency profiles - the anthropology degree and the engineering degree are far from each other because their profiles are different, whereas the mathematics degree is near the engineering degree because their profiles are similar. Distances between the points representing years are interpreted in the same way – each year point represents the profile of that year across the various disciplines.

This paper precisely proposes a link-analysis based technique allowing to discover relationships existing between elements of a relational database or, more generally, a graph. 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, 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 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.

Therefore, 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. In short, this paper has three main contributions:

- A two-step procedure for analyzing weighted graphs or relational databases is proposed
- It is shown that the suggested procedure extends correspondence analysis.
- A kernel version of the diffusion-map distance, applicable to directed graphs, is introduced.

II. Diffusion Map

A diffusion map is a machine learning algorithm for dealing with dimensionality reduction. The algorithm was first introduced by R.R. Coifman and S. Lafon in Applied and Computational Harmonic Analysis and Diffusion Maps and Geometric Harmonics. Unlike other dimensionality reduction methods such as Principle Component Analysis (PCA) and Multi-Dimensional Scaling (MDS), diffusion mapping is a non-linear method which focuses on discovering the underlying manifold in which the data is embedded. By integrating local similarities at different scales, the diffusion map gives a global description of the data set. Compared with other methods, diffusion maps are robust to noise perturbation and are computationally inexpensive.

Diffusion maps try to find a relation between distance) and probability. The basic observation is that if we take a random walk on the data, walking to a nearby data point is more likely than walking to another that is far away. Based on this, the connectivity between two data points, x and y, can be defined as the probability of walking from x to y in one step of the random walk. Usually, this probability will be defined as the kernel function on the two points, for example, the popular Gaussian kernel:

$$p(x,y) = k(x,y) = e^{-\frac{|x-y|^2}{\alpha}}$$

Here the kernel constitutes the prior definition of the local geometry of the data set. Since a given kernel will capture a specific feature of the data set, its choice should be guided by the application that one has in mind.

A. Diffusion Process

With the definition of connectivity between two points, we can define the diffusion matrix L (which is also a version of Laplacian matrix).

$$L_{i,j} = k(x_i, x_j)$$

Many works use the normalized matrix as

$$M = D^{-1}L$$

We can define the probability of moving from i to j in t steps as :

$$p(x_j, t|x_i) = M_{ij}^t$$

B. Diffusion Distance

The diffusion distance at time t between two points can be measured as the similarity of two points in the observation space with the connectivity between them. It is given by

$$D_t(x_i, x_j)^2 = \sum_y (p(y,t|x_i) - p(y,t|x_j))^2 w(y)$$

Here w(y) is a weighted function which can often be treated as w(y)=1

C. Diffusion Process

Now, we can give the definition of diffusion map as:

$$\Psi_t(x) = (\lambda_1^t \psi_1(x), \lambda_2^t \psi_2(x), \dots, \lambda_n^t \psi_n(x))$$

and it can be proven that

$$D_t^2(x_i, x_j) = ||\Psi_t(x_i) - \Psi_t(x_j)||^2$$

If we take the first k eigen vectors and eigen values, we get the diffusion map from the original data to a k dimension space which is embedded in the original space.

D. Diffusion Map Algorithm

The basic algorithm framework of a diffusion map is as:

- Step 1. Given the similarity matrix L
- Step 2. Form the normalized matrix $M = D^{-1}L$
- Step 3. Compute k largest eigenvectors and Eigen values of M^t
- Step 4. Use diffusion map to get the embedding matrix Y

E. A Kernel View of the Diffusion Map Distance

- The kernel version of the diffusion map is applicable to directed graphs while the original model is restricted to undirected graphs
- The extended model induces a valid kernel on a graph.
- The resulting matrix has the nice property of being symmetric positive definite—the spectral decompos -ition can thus be computed on a symmetric positive definite matrix, and finally
- 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.

This kernel-based technique will be referred to as the diffusion map kernel PCA or the KDM PCA.

III. Links between the Basic Diffusion Map and the Kernel Diffusion Map

While both representing the graph in a Euclidean space, where the nodes are exactly separated by the distances defined by (2), and thus providing exactly the same embedding, the mappings are, however, different for each method. Indeed, the coordinate system in the embedding space differs for each method.

In the case of the basic diffusion map, the eigenvector uk represents the kth coordinate of the nodes in the embedding space. However, in the case of the diffusion map kernel, since a kernel PCA is performed, the first coordinate axis corresponds instead to the direction of maximal variance in terms of diffusion map distance (2). Therefore, the coordinate system used by the diffusion map kernel is actually different than the one used by the basic diffusion map.

Putting the coordinate system in the directions of maximal variance, and thus computing a kernel PCA, is probably more natural. We now show that there is a close relationship between the two representations. We easily observe that the mapping provided by the basic diffusion map remains the same in function of the parameter t, up to a scaling of each coordinate/dimension (only the scaling changes). This is, in fact, not the case for the kernel diffusion map. In fact, the mapping provided by the diffusion map kernel tends to be the same as the one provided by the basic diffusion map for growing values of t in the case of an undirected graph. Indeed, it can be shown that the kernel matrix can be rewritten as $K_{DM} \propto U\Lambda^{2t}U^{T}$ where U contains the right eigenvectors of P; uk, as columns. In this case, when t is large, every additional dimension has a very small contribution in comparison with the previous ones.

IV. Stochastic Complementation

Although not always given an explicit name, a quantity which we shall refer to as a stochastic complement arises very naturally in the consideration of finite Markov chains. This concept has heretofore not been focused upon as an entity unto itself, and a detailed study of the properties of stochastic complementation has not yet been given. The purpose of the first part of this exposition is to explicitly publicize the utility of this concept and to present a more complete and unified discussion of the important properties of stochastic complementation.

The purpose of this section is to introduce the concept of a stochastic complement in an irreducible stochastic matrix and to develop some of the basic properties of stochastic complementation. These ideas will be the cornerstone for all subsequent discussions.

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 S1 and S2, respectively; usually $n_2 >> n_1$. Thus, the transition matrix repartitioned as:

$$A = \begin{bmatrix} O & A_{12} \\ A_{21} & O \end{bmatrix}.$$

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. Thus, performing a stochastic complementation allows focusing 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.

It can be shown that the matrix Pc is stochastic, that is, the sum of the elements of each row is equal to 1 ; it therefore corresponds to a valid transition matrix between states of interest. We will assume that this resulting stochastic matrix is a periodic and irreducible, that is, primitive . Indeed, Meyer showed in that if the initial chain is irreducible or a periodic, so is the reduced chain. Moreover, even if the initial chain is periodic, the reduced chain frequently becomes a periodic by stochastic complementation. 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, P_c, is primitive

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

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix};\ D = \begin{bmatrix} D_1 & O \\ O & D_2 \end{bmatrix};\ L = \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}$$

V. Analysing the Reduced markov Chain with Basic Diffusion Map: Link with Corresponcense Analysis

Once a reduced Markov chain containing only the nodes of interest has been obtained , only may want to visualize the graph in a low dimensional space preserving as accurately as 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.

A. Simple Correspondence Analysis

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 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.

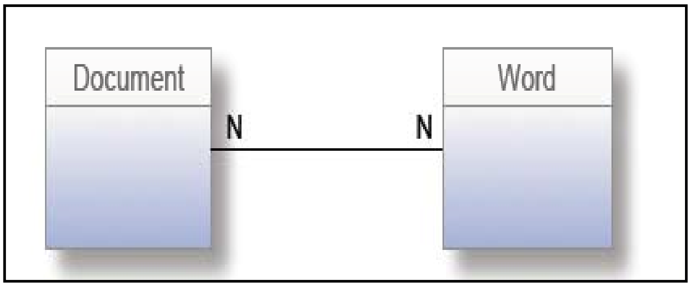


Fig. 1:

This situation can be modelled 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 associated n X n adjacency matrix and the corresponding transition matrix can be factorized as 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. The associated n X n adjacency matrix and the corresponding transition matrix can be factorized as

$$A = \begin{bmatrix} O & A_{12} \\ A_{21} & O \end{bmatrix},\ P = \begin{bmatrix} O & P_{12} \\ P_{21} & O \end{bmatrix},$$

Where O is a matrix full of zeroes

Suppose we are interested in studying the relationships between the attributes of the first variable x1 which corresponds to the n_1 first elements. By stochastic complementation (see Equation (10)), we easily obtain $P_C = P_{12}\,P_{21} = D_{11}A_{12}D_{21}A_{21}$. Computing the diffusion map for t=1 aims to extract the subdominant right-hand eigenvectors of P_c, which exactly corresponds to correspondence analysis. Moreover, it can easily be shown that Pc has only real non-negative 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 attribute explains 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.

B. Multiple Correspondence Analysis

Multiple correspondence analysis assigns a numerical score to each attribute of a set of p > 2 categorical variables. 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.

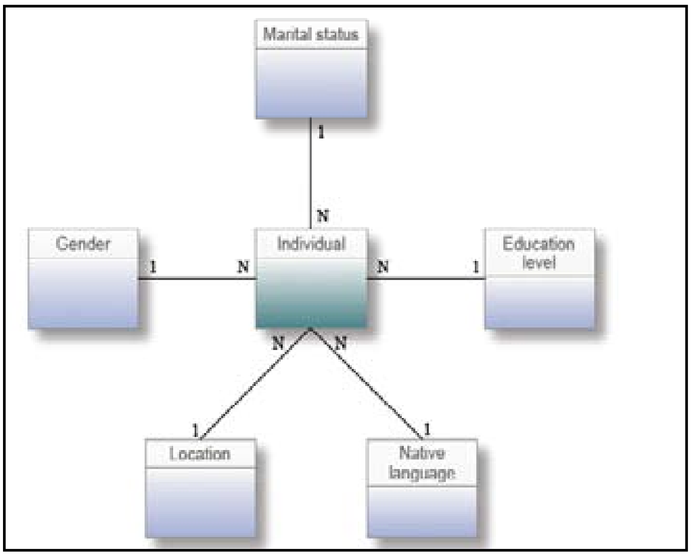


Fig. 2:

Let us first renumber the nodes in such a way that the attributes 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

$$A = \begin{bmatrix} O & A_{12} \\ A_{21} & O \end{bmatrix}.$$

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 are coded as indicator (dummy) variables, a row of the X matrix contains a 1 if the individual has the 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, therefore hiding the individuals nodes contained in the main table. By stochastic complementation (Equation (10)), the corresponding attribute-attribute transition matrix is

$$P_c = D_1^{-1}A_{12}D_2^{-1}A_{21} = \frac{1}{p}D_1^{-1}A_{12}A_{21}$$
$$= \frac{1}{p}D_1^{-1}X^TX = \frac{1}{p}D_1^{-1}F,$$

where the element of the frequency matrix $F = X^TX$, also called the Burt matrix, contains the number of co-occurences of the two attributes i and j, that is, the number of individuals having both attribute i and attribute j.

Thus, computing the eigen values and eigen-vectors of 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 with t = 1. If we are interested in the relationships between elements of the main table (the individuals) instead of the attributes, we obtain

$$P_e = \frac{1}{p}A_{21}D_1^{-1}A_{12} = \frac{1}{p}XD_1^{-1}X^T,$$

VI. Conclusion

A link analysis based technique allowing relationships existing in relationships existing in relational database. The database is viewed as a graph, where the nodes correspond to the elements contained in the tables and links correspond to the relation between the tables . a two step procedure is defined for analysing the relationships between elements of interest contained in a tables , 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 close related to correspondence Proposes to use stochastic complementation for extracting a subgraph containing the elements of interest from the original.

References

[1] C. D. Meyer,"Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems", SIAM Review, 31(2), pp. 240–272, 1989.

[2] 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.

[3] B. Nadler, S. Lafon, R. Coifman, I. Kevrekidis,"Diffusion maps, spectral clustering and eigen functions of Fokker-Planck operators", Advances in Neural Information Processing Systems 18, pp. 955–962, 2005.

[4] B. Nadler, S. Lafon, R. Coifman, I. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinate of dynamical systems", Applied and Computational Harmonic Analysis, 21, pp. 113– 127, 2006.

[5] C. Blake, E. Keogh, C. Merz,"UCI repository of machine learning databases", Irvine, CA: University of California, Department of Information and Computer Science, 1998.

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

[7] R.R. Coifman, S. Lafon,"Diffusion maps", Applied and Computational Harmonic Analysis, 21(1), pp. 5–30, 2006.

[8] A. Ihler,"Nonlinear Manifold Learning (MIT 6.454 Summary)", 2003.

[9] I.T. Jolliffe,"Principal component analysis", Springer- Verlag New York, 1986.
[10] S.S. Lafon,"Diffusion Maps and Geometric Harmonics", Ph.D thesis, Yale University, 2004.
[11] Joshua B. Tenenbaum, Vin de Silva, John c Langford,"A Global Geometric Framework".

Mr. G. Tatayyanaidu is a student of Akula Gopayya College of Engineering & Technology, Tadepalligudem. Presently he is pursuing his M.Tech (C.S.E) from this college and he received his B.Tech (C.S.E) from Bonam Venkata Chalamayya Institute Of Technology & Science (BVCITS), Batlapalem, JNTU Kakinada, East Godavari, Andhra Pradesh. His area of interest includes Data Structures, Object Oriented Programming, Operating Systems and Software Engineering.

Mr. K T V SUBBARAO is an excellent Associate Professor. He received M.Tech (CSIT) from Sagi Rama Krishnam Raju Engineering College, (SRKR) Bhimavaram, AndhraUniversity. Presently he is pursuing his P.hd in Computer Science & Engineering from Acharya Nagarjuna University. He is working as an Associate Professor in the Department of C.S.E, as Vice-Principal in Akula Gopayya College of Engineering and Technology. He has 13 years of teaching experience. He has published many papers in both National & International Journals. His area of Interest includes Data Communications & Networks, Data Warehouse and Data Mining, Database Management Systems and other advances in computer Applications. He designed and guided many Projects in the field of Computer Science and Information Technology.

Mr. M. Bala Krishna, well known Author and excellent Associate Professor. He received M.Tech (SE) from Godavari Institute of Engineering and Technology (GIET), Rajahmundry , JNTU Kakinada, East Godavari, Andhra Pradesh. He is working as Associate Professor and HOD, M.Tech Computer Science Engineering in Akula Gopayya College of Engineering and Technology. He has 09 years of teaching experience in various engineering colleges. To his credit he published a couple of publications both in National and International journals. His area of Interest includes Software Engineering, Information Security, Flavors of Unix Operating systems and other advances in computer Applications.