KDD-2002 Workshop Report
Fractals and Self-similarity in Data Mining: Issue and Approaches

Jafar Adibi
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
adibi@isi.edu

Christos Faloutsos
Carnegie Mellon University
Department of Computer Science
5000 Forbes Avenue
Pittsburgh, PA 15213-3891
christos@cs.cmu.edu

ABSTRACT

In this report we provide a summary of the first workshop on application of self-similarity and fractals in data mining: issues and approaches held in conjunction with ACM SIGKDD 2002, July 23 at Edmonton, Alberta, Canada.

Keywords

Data mining, Fractals, Self-similarity.

1. INTRODUCTION

Human brains, economic markets, network data, earthquake activity, music melodies and nature create enormously complex behavior that is much richer than the behavior of the individual component units. Self-similar stochastic processes, which introduced forty years ago, brought to the attention of probabilists and statisticians and they have been used in hydrology geophysics, biophysics, biology, financial economics and communications systems. Even recent measurements of new phenomena such as local-area and wide-area traffic, web site traffics and agents' behavior have shown self-similarity at a wide range of time scales.

Recently, several techniques have been proposed to apply data mining techniques through fractal dimensions and self-similar characteristics in different domains. These techniques include but not limited to: dimension reduction, predictive modeling, using self-similar characteristics to mine databases such as association rules, clustering, classification, modeling and outlier discovery, selectivity estimation, spatial databases, R-trees, Quadtrees and model distributions of data. Although fractals and self-similarities in Data Mining have attracted researchers' attention, this is still an open research issue. We believe that much progress could be achieved from a concerted effort and a greater amount of interactions between researchers in different research areas.

The purpose of this workshop was to provide a forum to foster such interactions, discuss the new achievements and identify future research directions. Hence, to high light these efforts we organized the first workshop on application of self-similarity and fractals in data mining held in conjunction with ACM SIGKDD 2002.

The program of the workshop included of introductory talk, an invited talk, six contributed papers and a panel at the end of the program. The on-line proceedings for the invited papers are available at: www.isi.edu/~adibi/FratalKDD02.htm.

The introduction by Christos Faloutsos (CMU) gave the definition of fractals, some examples like the Sierpinski triangle, and made the observation that fractals and power laws are closely related and at the end he opened the floor for the rest of the program.

2. INVITED TALK

In his invited talk Jon Kleiberg (Cornell) talked about the Information Networks, Power Laws, and Small-World Phenomena. On the information network he stated that if we were to generate a random graph on n nodes by including each possible edge independently with some probability p, then the fraction of nodes with d neighbors would decrease exponentially in d. But for many large networks including the Web, the Internet, collaboration networks, and semantic networks it quickly became clear that the fraction of nodes with d neighbors decreases only polynomially in d; to put it differently, the distribution of degrees obeys a power law. What processes are capable of generating such power laws, and why should they be ubiquitous in large networks? The investigation of these questions suggests that power laws are just one reflection of the local and global processes driving the evolution of these networks.

He continued his talk by discussion of link analysis with the goal of searching for and ranking high-quality content. A lot of the methods to do this, however, uncover richer structure in the network - densely connected communities focused on a particular topic. The problem of uncovering such community structure differs in some interesting ways from the heavily studied problems of clustering and graph partitioning rather than trying to decompose the full network into a few pieces, the goal is to identify small, densely connected regions that even put together may not account for much of the whole graph.

The talk continued with overview on the amazing small-world phenomena, the principle that we are all linked by short chains of acquaintances. This observation was inaugurated as an area of experimental study in the social sciences through the pioneering work of Stanley Milgram in the 1960's. He showed that no decentralized algorithm, operating with local information only, can construct short paths in these networks with non-negligible probability. He defined an infinite family of network models that

naturally generalizes the Watts-Strogatz model, and showed that for one of these models, there is a decentralized algorithm capable of finding short paths with high probability. In general he provided a strong characterization of this family of network models, showing that there is in fact a unique model within the family for which decentralized algorithms are effective. He also mentioned that small-world navigation model to power-law distributions in which nodes know the degree of their neighbors; affects the sub-linear search time and facilitates the simultaneous search. The talked wrapped up by overview on open issues and possible future work in this field.

3. CONTRIBUTED PAPERS

The contributed papers spanned a series of interesting topics which is listed as following.

3.1 Fractals and Self-Similarity for Advance Application

Jafar Adibi (USC/ISI), Wei-Min Shen (USC/ISI) and Eaman Noorbakhsh (CALTECH/JPL) provided a review on fractals and self-similarity for advance application including Agent Theory, Self-Similar Layered Hidden Markov Model and Link Discovery.

Jafar Adibi argued that a long lasting question, still in debate among multi-agent system researchers, is how to quantitatively measure the amount of "teamwork" in a multi-agent system. Previous attempts are mostly qualitative and in some cases ad hoc, and have not provided any satisfactory solutions. In his talk he provided preliminary results of his novel technique to measure the fractal dimension of a multi-agent system, and showed that such a measurement is directly related to the amount of teamwork in the system. He also presented his observation of RoboCup simulation league (agents playing soccer in a simulated environments) that despite the complicated behaviors of the synthetic agents in this dynamic domain, agents' behaviors are self-similar (i.e., macro and micro structures are reflected in each other) and the fractal measurements in their behaviors are directly related to the team strategies in the competitions.

He also introduced Self-Similar Layered HMM and illustrated that it is a better estimation than flat Hidden Markov Model when data shows self-similar property and exploit this property to reduce the complexity of model construction. . He showed how the embedded knowledge of self-similar structure can be used to reduce the complexity of learning and increase the accuracy of the learned model. In addition he explained three different types of self-similarity along with some result on synthetic data and experiments on Network data

The last application introduced by Jafar Adibi was his new observation in Link Discovery problem. He illustrated that the distribution of the number of links/info/activities associate with agents in a real world database environments shows ZIPF law's distribution. This information could be employed to answer different questions in link discovery problem such as classification, clustering and object consolidation.

3.2 WEKWEK : A Study in Fractal Dimension and Dimensionality Reduction

Many real-life datasets have a large number of features some of which are often highly correlated. This is referred to as the curse of dimensionality in data mining literautes. Elena Eneva (CMU), Krishna Kumaraswamy (CMU) and Matteo Matteucci (CMU) investigated the relationship among several dimensionality reduction methods and the intrinsic dimensionality of the data in the reduced space, as estimated by the fractal dimension. They showed that a successful dimensionality reduction/feature extraction algorithm projects the data into a feature space with the dimensionality close to the intrinsic dimensionality of the data in the original space and preserves topological properties.

In addition, they provided a comparative study between the dimensionality reduction techniques using fractal dimension and other well-known techniques such as principle component analysis (PCA), factor analysis (FA) and self-supervised multi layer perceptron.

3.3 Accelerating Clustering methods through Fractal based Analysis

Changhao Jiang (CMU), Yiheng Li (CMU), Minglong Shao (CMU) and Peng Jia (CMU) attacked the problem of clustering of large high dimensional datasets. They presented a new fractal-based sampling tool to accelerate prevailing clustering methods in two aspects: 1) as a preprocessing step to iteratively sample the original dataset into critical-size of dataset; 2) advising clustering methods with critical input parameters which is specific to different methods.

The key idea of their work is based on the assumption that id a samples subset has similar "self-plot" to that of original dataset, then it preserve original datasets' distribution pattern and can be used in place of original dataset for clustering. With this assumption a fractal-based sampling tool is added as a preprocessing step before clustering. A given large high dimensional dataset first goes through this sampling tool and is reduced to critical sized subset is fed into a clustering method. The preprocessing step reduces dataset's size and it advises successive clustering method with some critical input parameters which are obtained through "self-plot" analysis. Their experimental result of applying such tool with BIRCH clustering method suggests the tool's effectiveness and applicability in both aspects. In addition they illustrated their experiments on OPTICS and CURE clustering techniques in which both result were positive.

3.4 SELFIS: A Tool for Self-Similarity and Long Range Dependence Analysis

Over the last few years, the network community has started to rely heavily on the use of novel concepts such as fractals, self-similarity, long-range dependence and power-laws. Especially evidence of fractals, self-similarity and long-range dependence in network traffic have been widely observed. Despite their wide use, there is still much confusion regarding the identification of such phenomena in real network traffic data. For one, the Hurst exponent can not be calculated in a definitive way, it can only be estimated. Second, there are several different methods to estimate the Hurst exponent, but they often produce conflicting estimates. It is not clear which of the estimators provides the most accurate estimation. In this extended abstract, we make a first step towards a systematic approach in estimating self-similarity and long-range dependence.

Thomas Karagiannis (UCR) and Michalis Faloutsos (UCR) presented a java based tool called SELFIS that will automate the

self-similarity analysis. SELFIS is the first attempt to create a stand-alone, free, open-source platform to facilitate self-similarity analysis. They showed the use of SELFIS and described the methodologies that currently incorporate in real Internet data. In addition, they presented an intuitive approach to validate the existence of long-range dependence.

SELFIS facilitates practitioners by providing practical ways for long-range dependence estimation and it can become a reference point in estimating long-range dependence in time-series. A number of different estimators that capture various features of long-range dependent behavior is provided by the software.

3.5 How to Use Fractal Dimension to Find Correlations between Attributes

One of the challenges when dealing with multimedia information, which usually is massive and composed of multidimensional data, is how to index, cluster and retain only the relevant information among the items of a large database. In order to reduce the information to the meaningful data, techniques such as attribute selection have been used. Some of the well-known dimensionality reduction techniques are the principal component analysis (PCA) and the singular value decomposition (SVD). These techniques either map the data sets to other spaces or generate a set of additional axes. However, this is not quite attribute selection.

Elaine Parros Machado de Sousa (University of Sao Paulo), Caetano Traina Jr. (University of Sao Paulo), Agma J. M. Traina (University of Sao Paulo) and Christos Faloutsos (CMU) presented a novel technique to discover groups of correlated attributes in a data set, using the fractal dimension concept. They also presented an algorithm to find correlation groups that is linear on the number of attributes and on the number of elements in the data set. These techniques represents an important achievement on finding information about a data set, as it can discover if there is more than one group of attributes that establishes the correlations in the data set, and also identify the attributes that pertain to each group. The result of such technique can be employed (1) to find a subset of the attributes that, besides concisely representing the data, can be used when creating indexes or applying data mining techniques over the data without compromising the results; (2) to establish which attributes are correlated between them.

3.6 Neural Network Modeling of Self-Similar Teletraffic Patterns

The significant characteristic of bursty traffic is self-similarity. Self-similarity is the main reason for observing so called burst within burst patterns across a wide range of time scales as one of the unique characteristics of nonlinear systems with fractal nature. Perceptron neural networks, because of their nonlinear nature and simple method of learning, provide a powerful tool to model bursty traffic patterns.

Homayoun Yousefi'zadeh (UCI) studied modeling of self-similar traffic patterns with a multi-purpose fixed structure system, i.e, a neural network. He investigated perceptron neural networks and back propagation learning algorithm for the task of modeling. He used a fixed structure neural network with three hidden layers in order to model source- and aggregated-level traffic patterns. The neural network had eight neurons in its input layer, and twenty neurons in each of its three hidden layers. The number of neurons in its output layer was one. He applied eight consecutive samples of the pattern as the input of the network and used the ninth sample as the desired output in each iteration of the learning algorithm. He utilized neural networks in modeling artificially generated self-similar traffic patterns by intermittency family of chaotic maps. The number of samples required for the training of the fixed structure neural network depended on the steady state behavior of the traffic pattern and the network load. For the learning phase he made use of enhanced back propagation learning algorithm to shorten the learning phase after which the network was able to predict the sequential samples of the corresponding traffic pattern within an error bound. At the end he followed each recalling phase with a training phase to be able to cope with the divergent behavior of the trained neural network and provided some intuitive reasoning relying on sensitivity to the variations of initial conditions argument discussed in chaos theory to explain the observed divergent behavior

4. ACKNOWLEDGMENTS

We would like to thank workshop invited speaker Prof. Jon Kleinberg, authors, panelists and participants for contributing to the success of the workshop. Special thanks are due to the program committee for their support and help. We are also thankful to the members of KDD workshop committee for providing the opportunity to stage this event.

About Authors:

Jafar Adibi is a Research Associate at the University of Southern California's Information Sciences Institute. He has published over 20 refereed articles, achieved 3 international medals, and he holds two pending patents. His research interests include data mining for pattern recognition and link discovery, fractals and self-similarity for advance application, and machine learning.

Christos Faloutsos is a Professor at Carnegie Mellon University. He has received the Presidential Young Investigator Award by the National Science Foundation (1989), four ``best paper'' awards (SIGMOD 94, VLDB 97, KDD01 runner-up, Performance02 best student paper), and four teaching awards. He is a member of the executive committee of SIGKDD; he has published over 120 refereed articles, one monograph, and holds four patents. His research interests include data mining for streams and networks, fractals, indexing methods for spatial and multimedia bases, and data base performance.

Program Committee