DESCRIPTION of the DATASETS

Overview

In our research we used the following input files: CRA_W_DMP_dataset_N.xml, OSCAR_N.xml, toygraph_N.xml, guass_centroid_N[i].xml, written in a XGMML format, loaded by Prefuse.

Abstract

The meaningful data in the datasets represents a variety of forms including continuous numerical values, discrete values, symbolic and categorical values. These datasets allow to automatically select different subsets of variables. The datasets were created to allow to determine the nature of the relationships between variables in the *n-dimensional* space.

Data Structure Represented in the Datasets

Each dataset has similar structure. All of them represent a directed graph. Each dataset contains a different amount of nodes. For each node in the graph there is an ID number, which has no influence on how the program interprets the node; the ID numbers don't have to be in order. Each node also includes a Label, which has a string assigned to it and is seen on the graph layout. Each node can contain an infinite amount of attributes with an assigned attribute name and a value, which can be in a variety of forms.

In the demo GraphLayoutMAPLE_N.java, the following were used for standardization of the interpretation of the variety of data's form from above named datasets:

-for image - attribute names with “image”

-for categorical data – attribute names begin with symbol “@”

-for discrete data – attribute names begin with symbol “#”

-for text strings – attribute names begin from “$”

-for numerical values – any attribute name that cannot begin with the listed above attribute names.

After listing all of the nodes, the datasets include the list for all of the edge connections between nodes. Each edge connection includes a source number, which represents the ID number of the node, and a target number, which represents the ID number of another node. Each source number has an outgoing edge, and the target – an ingoing edge. Each edge includes a weight property, the meaning of which is different for every dataset.

A commented description of each node set and edge set is also included in the datasets for a better understanding of the data shown on the graph.

Description of “toygraph_N.xml”

Dataset “toygraph_N.xml” was created by using the package data_generation_MAPLE_N from the console with randomly connected edges.

Represented dataset “toygraph_N.xml” includes 45 nodes and 110 edges.

Each node creates a four-dimensional space.

ID numbers are in order, each label is n[i] where i is in the same order as id number

Each node contains the following attribute names: north, south, east, west; which were prompted from console.

For each attribute name the allowed values that created the vector were prompted.

Values for the nodes chosen were generated randomly by the package from the continuous numerical values in the vector. The range of values depends on the prompted data.

The edge connection was created randomly by the program WriteXML_N.java from the package edu.umbc.cs.maple.data_generation_MAPLE_N.

The weight of each edge is set by the package to “1”.

Detailed description of the creation of “toygraph_N.xml” is in the package edu.umbc.cs.maple.data_generation_MAPLE_N.

PROBLEMS: none.

Description of “OSCAR_N.xml”

Dataset “OSCAR_N.xml” was created by hand using the Official Academy Awards Database from http://www.oscars.org/awardsdatabase/index.html

Represented dataset “OSCAR_N.xml” includes 56 nodes and 168 edges.

Different nodes represent different dimensional space.

ID numbers are in order.

The first node (id=”1” label=”OSCAR”) includes attributes' names and categorical values, which represent the year when the ceremony for the Academy Awards was held. The range of values is from 1999-2004.

Each of the node label from id=”2” to id=”50” contains:

- a string with the first name initial and the last name of the actor/actress.
- pictures of actors/actresses which were collected from different websites.
- award year when the actor/actress took part in the awards ceremony and with the corresponding values of the year of the awards ceremony.
- gender with categorical values 1 – actress; 2 – actor.
- total number of nomination which represents the number of events the actor attended in the academy awards. Range of values is from 1 to 13.
- film, which was presented at the academy awards.
- if an actor/actress won an Oscar, that information is noted and the values represent how many Oscars the person won.
- These nodes can contain additional information. For example, if an actor/actresses took part in the awards ceremony during 1999-2004, or was in more than one movie. This information is included in the node.

Nodes with id=51 to 56 represent the number of academy awards, which are discrete values. The attributes of these nodes contain the year when the academy awards was held.

The edge connection can visually be divided into subsets:

- Actors/Actresses which have won the Oscar are connected to the Oscar node

- Connection to the actor/actress with the corresponding academy awards.
- Connection between the actor/actress, which won the Oscar in one particular academy awards.
- The weight of the edges represents the number of Oscars statues an actor/actress has won.

PROBLEMS: A problem with this dataset appeared when one of the servers, which the image of an actor/actress was on, did not allow to externally show an image. Loading of the images requires a few seconds.

Description of "CRA_W_DMP_dataset_N.xml"

Dataset "**CRA_W_DMP_dataset_N.xml**" was created by hand using the data from the official site "Mentoring Undergraduate Women in Computing Research CRA-W Distributed Mentor Project (DMP) Summer 2005 Awards" http://www.cra.org/Activities/craw/dmp/awards/2005/2005.php and data from Academic Ranking of World Universities 2004/Top 500 World Universities from the following website http://ed.sjtu.edu.cn/rank/2004/top500list.htm

Represented dataset "**CRA_W_DMP_dataset_N.xml**" includes 107 nodes and 255 edges.

Different nodes represent different dimensional space.

ID numbers are not in order.

The first node (id="1" label="CRA-W DMP") includes attribute name image which is a modified logo of the DMP.

Each of the node label from id="2" to id="45" represent the name of the institution that participated in the DMP. These node contains following attributes names:

- @DMP – to represent the categories institution by type of participants The value of the type of the participants are categorical value in the range [1-3]:
  1 – for representing student's institutions ;
  2 – for mentor's institutions;
  3 – for institution from witch students and mentors participated in the DMP.
- @status represents type of the institution with categorical values in the range[1-4]
  1 is seated to categorize University;
  2- for Colleges;
  3 – for Institutes;
  4 – for Research Centers;
- WR - to represents the world rank of the institution with corresponding value from Academic Ranking of World Universities. Range [1-500].
- Total student# - the total number of students that participate in DMP in the named institution . Range [0-9].
- Total mentor# - the total number of mentor that participate in DMP in the named institution . Range [1-5].

Each of the node label from id="46" to id="115" represents the name of the student or mentor that participated in the DMP. These node contains following attribute names:

- @participant to represent type of participated person in the DMP with the categorical values 1- for student; 2 – for mentor.
- @Gender with categorical values 1 – for women. 2 – for men

The node label from id=”89” to id=”115” contain additional attributes names

- student# - total number of students that joined to the mentor.
- pictures of mentors which were collected from mentor's and student's websites.

The edge connection can visually be divided into subsets:

- Edges DMP_M_univer represent the connection between DMP and the mentors' Universities which are currently participating in the internship program. The weight represents the number of mentors for the current University.
- Edges DMP_S_univer represent the connection between DMP and the students' Universities which are currently participating in the internship program. The weight represents the number of students from the current University.
- Edges s-univer-student and edges unvier-student-s represent the connection between the student's university and the student. Weight = 1.
- Edges m-univer-mentor and edges univer-m-mentor represent the connection between the mentor's university and the mentor. Weight represents how many students are currently working for the mentor
- Edges mentor-student and edges student-mentor represent the connection between the student and the mentor. Weight = 1.

Description of "guass_centroid_N[i].xml"

Datasets "guass_centroid_N[i].xml" were created to produce clusters with meaningful properties. The created clusters in the datasets were distributed according to multivariate Gaussian properties. Every dataset represents a directed graph. The datasets contain nodes in the range [50 -200] plus centroids that are in the range [2-4]. The nodes in the different datasets represent different dimensional space in the range [4-5]. All nodes have following structure:

- id numbers are in order (total number of id depends of the amount of the nodes in the dataset plus amount of the clusters);
- each id has corresponding label:

a)the label c[i] represents the cluster's node, where i is the number of the cluster in the dataset;

b) label c[i]n[j], where the node is represented by the "i" cluster and j – the ordered number of nodes in the cluster [i];

- each node contains attributes names x[k], where k represents the k-dimensional space and numerical value in the range [0-100]. The value for each centroid was chosen randomly from an M-vector of cluster means that were created using multi-variance Gaussian distribution by Matlab. The values for the nodes which are not centroids were chosen from an M-vector of cluster variances using covariance MxM matrix with a constance variance by Matlab.

After listing all of the nodes, the datasets include the list for all of the edge connections between nodes. The edges were added between a pair of nodes with a probability which depends on whether the nodes are in the same cluster:

- the probability of node connection is 0.1 for connection the edges in the same cluster;
- the probability of node connections between different clusters is 0.05.

The edge connection was created randomly by the program WriteXML_N.java from the package edu.umbc.cs.maple.data_generation_MAPLE_N.
The weight of each edge is set by the package to "1".

The final structure of each datasets was created by hand joining list of the nodes created by the Matlab and edges connections that were created by above mention package.

Author of the datasets:

Nataliya Lozova
LSU student
nlozov1@lsu.edu
Summer DMP Internship Program 2005 at UMBC
August 1, 2005
Revised: August 3, 2005
Revised: August 9, 2005