Module: CLUTO Toolkit

Draft: 10/21/2010

1) Module Name

CLUTO Toolkit

2) Scope

The module briefly introduces the basic concepts of Clustering. The primary focus of the module is to describe the usage of CLUTO, a clustering Toolkit, comprised of various algorithms.

3) Learning Objectives

At the end of the module, a student will be able to

i. Explain the basic concepts of Clustering and its relevance in digital libraries.

ii. Explain the different types of clustering algorithms.

iii. Use the CLUTO clustering software package and run programs to cluster objects using CLUTO.

4) 5S Characteristics of the Module

i. **Streams:** Input stream includes an input file containing information about objects to be clustered and their dimensions, clustering program name, and the number of clusters required. Output stream is a file wherein objects are grouped into desired number of clusters using a criterion function.

ii. **Structures:** Input data is stored either as a matrix file or graph file and output data is stored as a tree file or clustering solution file.

iii. **Spaces:** Feature spaces of objects are used by the vcluster program. The scluster program operates on the similarity space between the objects**.**

iv. **Society:** End users of the system are those who work in the fields of digital libraries, genetics, bioinformatics, biochemistry, customer services, etc.

v. **Scenarios:** Situations where users submit a set of documents and a similarity criterion to find clusters of related documents. For e.g., grouping a set of documents (objects) based on the terms (dimensions) they contain.

5) Level of Effort Required:

i. Prior to Class: 2 hours of reading

ii. In Class: 3 hours

a. 2 hours for learning the basics of clustering and implementation of CLUTO

b. 1 hour for class discussions, exercises, and activities

6) Relationship with other modules:

Close connections with:

i. *R Project for statistical computing module*

R-Project module is used for visualization and statistical analysis. Similarly, the CLUTO module is used to display statistical information about the clusters and display cluster, matrix graphs and tree plots using visualization.

ii. *Weka-3 module*

Weka-3 module contains tools for data preprocessing, classification, clustering and visualization. Most of these features are provided by CLUTO as well.

7) Prerequisite knowledge required:

Basic knowledge of UNIX commands, statistics and probability.

8) Introductory Remedial Instruction:

None

9) Body of Knowledge

i. What is clustering?

a. Clustering algorithms group a set of documents into subsets or clusters using a similarity or criterion function.

b. Documents within a cluster should be as similar as possible, i.e., the intra-cluster similarity should be high.

c. Documents in one cluster should be as dissimilar as possible from documents in other clusters, i.e., the inter-cluster similarity should be low.

d. Clustering can be classified into:

i. Flat Clustering and Hierarchical Clustering

a. Flat clustering creates a flat set of clusters without any explicit structure that would relate clusters to each other.

b. Hierarchical clustering finds successive clusters using previously established clusters. These algorithms usually are either agglomerative ("bottom-up") or divisive ("top-down").

ii. Hard Clustering and Soft Clustering

a. Hard clustering computes a hard assignment – each document is a member of exactly one cluster.

b. The assignment of soft clustering algorithms is soft – a document's assignment is a distribution over all clusters. In a soft assignment, a document has fractional membership in several clusters.

ii. What is CLUTO?

a. CLUTO is a software package for clustering low and high dimensional datasets and for analyzing the characteristics of the various clusters.

b. CLUTO uses clustering algorithms to divide data into clusters such that intra-cluster similarity is high and inter-cluster similarity is low.

c. CLUTO provides three different classes of clustering algorithms that operate either directly in the object's feature space or in the similarity space.

d. Algorithms are based on the partitional, agglomerative, and graph partitioning paradigms.

e. CLUTO provides a total of seven different criterion functions.

f. CLUTO provides tools for analyzing the discovered clusters to understand the relations between the objects assigned to each cluster and the relations between the different clusters.

iii. CLUTO Programs

a. Types of CLUTO Programs:

CLUTO consists of two stand-alone programs:

i. vCluster: 'V' stands for vector. This program takes actual multidimensional representation of the objects as the input.

ii. sCluster: ‘S’ stands for similarity. This program takes a similarity matrix of the objects as the input.

b. Calling Sequence:

CLUTO programs are invoked using the following calling sequence:

Program-name [optional parameters] file-name NClusters

The calling sequence consists of:

i. Program-name: vcluster or scluster.

ii. Two required parameters: These parameters include a) file-name – an input file containing information about objects and their dimensions b) NClusters - the number of clusters required.

iii. Additional optional parameters: These parameters are used to : a) control aspects of the clustering algorithm b) control type of analysis and reporting of clusters and c) control visualization of clusters.

c. Statistics:

CLUTO programs can be used to display statistical information about:

i. Clustering solution.

ii. Time taken to perform clustering.

iv. Criterion Functions

A key feature of a CLUTO clustering algorithms is that the problem is treated as an optimization process which is to maximize or minimize a criterion function. Criterion functions can be broadly classified into four categories: internal, external, hybrid, and graph based criterion functions. These are described and analyzed in [11]. Six criterion functions used for document clustering are described in [6].

CLUTO provides a total of seven different criterion functions that can be used to drive both partitional and agglomerative clustering algorithms. These are described in [2, 10, and 11].

Criterion Function	Optimazition Function	
$\mathcal{I}_1$	maximize $\sum_{i=1}^{k} \frac{1}{n_i} \left(\sum_{v,u \in S_i} \text{sim}(v,u) \right)$	(1)
$\mathcal{I}_2$	maximize $\sum_{i=1}^{k} \sqrt{\sum_{v,u \in S_i} \text{sim}(v,u)}$	(2)
$\mathcal{E}_1$	minimize $\sum_{i=1}^{k} n_i \frac{\sum_{v \in S_i, u \in S} \text{sim}(v,u)}{\sqrt{\sum_{v,u \in S_i} \text{sim}(v,u)}}$	(3)
$\mathcal{G}_1$	minimize $\sum_{i=1}^{k} \frac{\sum_{v \in S_i, u \in S} \text{sim}(v,u)}{\sum_{v,u \in S_i} \text{sim}(v,u)}$	(4)
$\mathcal{G}'_1$	minimize $\sum_{i=1}^{k} n_i^2 \frac{\sum_{v \in S_i, u \in S} \text{sim}(v,u)}{\sum_{v,u \in S_i} \text{sim}(v,u)}$	(5)
$\mathcal{H}_1$	maximize $\frac{\mathcal{I}_1}{\mathcal{E}_1}$	(6)
$\mathcal{H}_2$	maximize $\frac{\mathcal{I}_2}{\mathcal{E}_1}$	(7)

Table 1: The mathematical definition of CLUTO's clustering criterion functions. The notation in these equations are as follows: k is the total number of clusters, S is the total objects to be clustered, S_i is the set of objects assigned to the ith cluster, n_i is the number of objects in the ith cluster, v and u represent two objects, and $\text{sim}(v,u)$ is the similarity between two objects.

Table adapted from CLUTO manual [2]

v. CLUTO Algorithms

Clustering algorithms used in CLUTO depend on:

a. Cluster Types:

i. Cluster types are based on similarity-view between the clusters, i.e., the relations between a) cluster objects and b) dimensions of feature space.

ii. Clusters are of two types:

a. **Globular clusters:** The objects in these clusters have high pair-wise similarities with dense subspaces.

b. **Transitive clusters:** The objects in these clusters have low pair-wise similarities. Each cluster consists of several sub-clusters connected by strong paths with high similarity edges.

b. Similarity Measures:

The similarity measures used by CLUTO algorithms are:

i. Cosine and correlation coefficient measures: The object similarity is based on the direction of the corresponding vectors but not the magnitude.

ii. Euclidean distance: The object similarity is based on both direction and magnitude of the corresponding vectors.

iii. Jaccard coefficient: The object similarity is based on angle and magnitude of the corresponding vectors.

c. Scalability of Algorithms:

The scalability of CLUTO algorithms depends on:

i. Space complexity

ii. Time complexity

vi. History of CLUTO

a. CLUTO was developed at Karypis Lab, University of Minnesota Twin Cities.

b. The first change to CLUTO, ver.: 1.5 was released on 01/08/2002. Features of agglomerative clustering algorithms, cluster visualization capability, dense input file support were added.

c. The latest version of CLUTO, ver.: 2.1.2a was released on 01/09/2007. It included a build for Windows X86_64.

d. Details about the specific features of the different releases can be found at:

http://glaros.dtc.umn.edu/gkhome/cluto/cluto/changes

vii. CLUTO Schematic Diagram

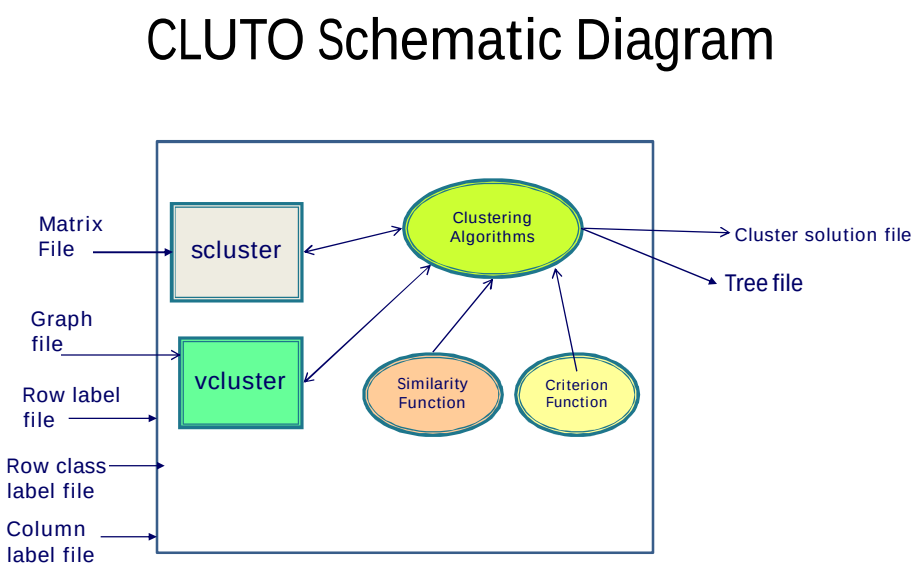


Figure 1: CLUTO schematic diagram

viii. Example application areas of CLUTO

a. Digital Libraries - To cluster documents (objects) based on the terms (dimensions) they contain.

b. Customer Services - Amazon.com may group customers (objects) based on the types of products (nooks, music products - dimensions) they purchase, etc.

c. Genetics - To cluster genes (objects) based on their expression levels (dimensions)

d. Biochemistry - To cluster proteins (objects) based on the motifs (dimensions) they contain.

ix. Features of CLUTO

a. Multiple classes of clustering algorithms: partitional, agglomerative, & graph-partitioning based.

b. Multiple similarity/distance functions: Euclidean distance, cosine, correlation coefficient, extended Jaccard, user-defined.

c. Numerous novel clustering criterion functions and agglomerative merging schemes.

d. Traditional agglomerative merging schemes: single-link, complete-link, UPGMA.

e. Extensive cluster visualization capabilities and output options: postscript, SVG, gif, xfig, etc.

f. Multiple methods for effectively summarizing the clusters: most descriptive and discriminating dimensions, cliques, and frequent item sets.

g. Can scale to very large datasets containing hundreds of thousands of objects and tens of thousands of dimensions.

x. Relation to IR Concepts

CLUTO uses the following IR concepts:

a. Similarity functions: Jaccard Coefficient, cosine similarity measure, Euclidean distance.

b. Cluster pruning techniques.

c. Flat clustering techniques.

d. Hierarchical clustering techniques.

xi. Input file formats in CLUTO

Input files used in CLUTO are of two types:

a. Matrix Format

i. This is the primary input for CLUTO's vcluster program.

ii. Each row of this matrix represents a single object.

iii. Columns correspond to the dimensions (i.e., features) of the objects.

iv. Matrix files are classified into two types:

a. Dense Matrix Format

- The first line of the matrix file contains exactly two numbers, both of which are integers. The first integer is the number of rows in the matrix (n) and the second integer is the number of columns in the matrix (m).

- Each subsequent line contains exactly m space-separated floating point values, such that the ith value corresponds to the ith column of A.

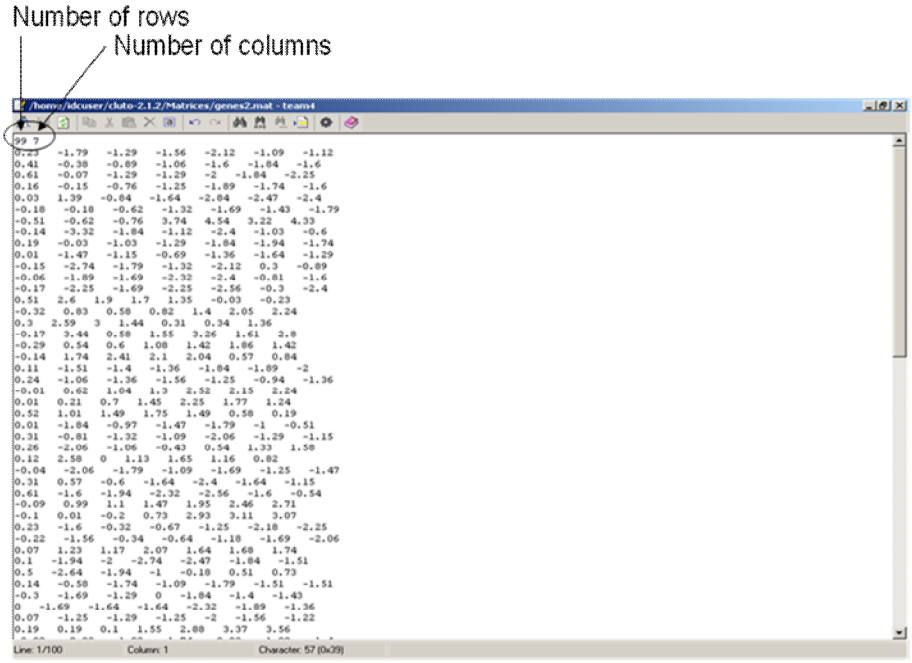


Figure 2: Dense Matrix file format

b. Sparse Matrix Format

- The first line contains information about the size of the matrix, while the remaining n lines contain information for each row of A. In CLUTO's sparse matrix format only the non-zero entries of the matrix are stored.

- The first line of the matrix file contains exactly three numbers, all of which are integers.

- The first integer is the number of rows in the matrix (n), the second integer is the number of columns in the matrix (m), and the third integer is the total number of non-zeros entries in the n × m matrix.

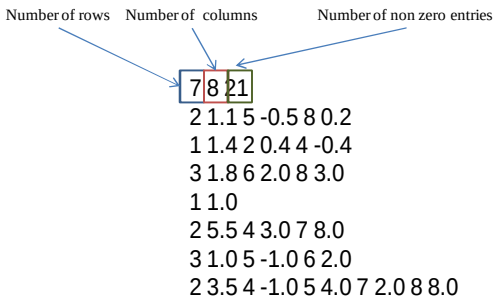


Figure 3: Sparse Matrix file format

b. Graph Format

i. This is the primary input for CLUTO's vcluster program. It is a square matrix.

ii. It specifies the similarity between the objects to be clustered.

iii. A value at the (i, j) location of this matrix indicates the similarity between the ith and the jth object.

iv. Graph files are classified into two types:

a) Dense Graph Format

- The first line of the file contains exactly one number, which is the number of vertices n of the graph.
- The remaining *n* lines store the values of the *n* columns of the adjacency matrix for each one of the vertices.
- Each line contains exactly n space-separated floating point values, such that the ith value corresponds to the similarity to the vertex equaling the row number of the current row of the graph.

b) Sparse Graph Format

- The first line of the file contains exactly two numbers, both of which are integers. The first integer is the number of vertices in the graph (n) and the second integer is the number of edges in the graph.
- The (i + 1)st line of the file contains information about the adjacency structure of the ith vertex.
- The adjacency structure of each vertex is specified as a space-separated list of pairs. Each pair contains the number of the adjacent vertex followed by the similarity of the corresponding edge.

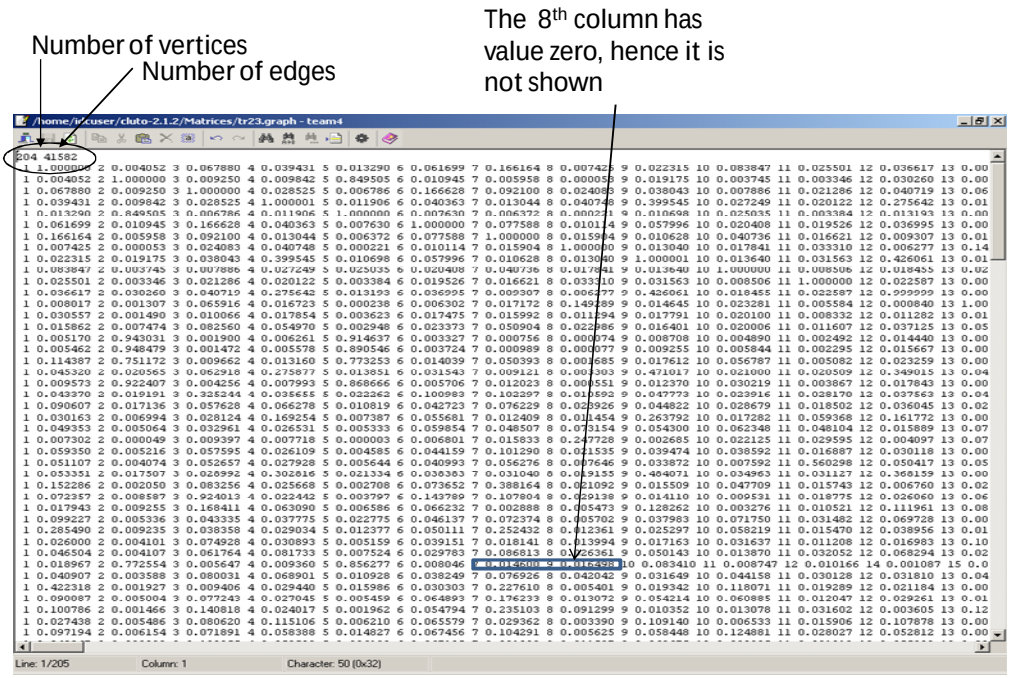


Figure 4: Sparse Graph file format

xii. Output file formats in CLUTO

Output files used in CLUTO are of two types:

a. **Clustering File Solution**

- The clustering file of a matrix with n rows consists of n lines with a single number per line. The ith line of the file contains the cluster

number that the ith object/row/vertex belongs to. Cluster numbers run from zero to the number of clusters minus one.

b. **Tree File**

- The tree produced by performing a hierarchical agglomerative clustering on top of the *k*-way clustering solution produced by vcluster is stored in a file in the form of a *parent* array.

- The ith line contains the parent of the ith node of the tree.

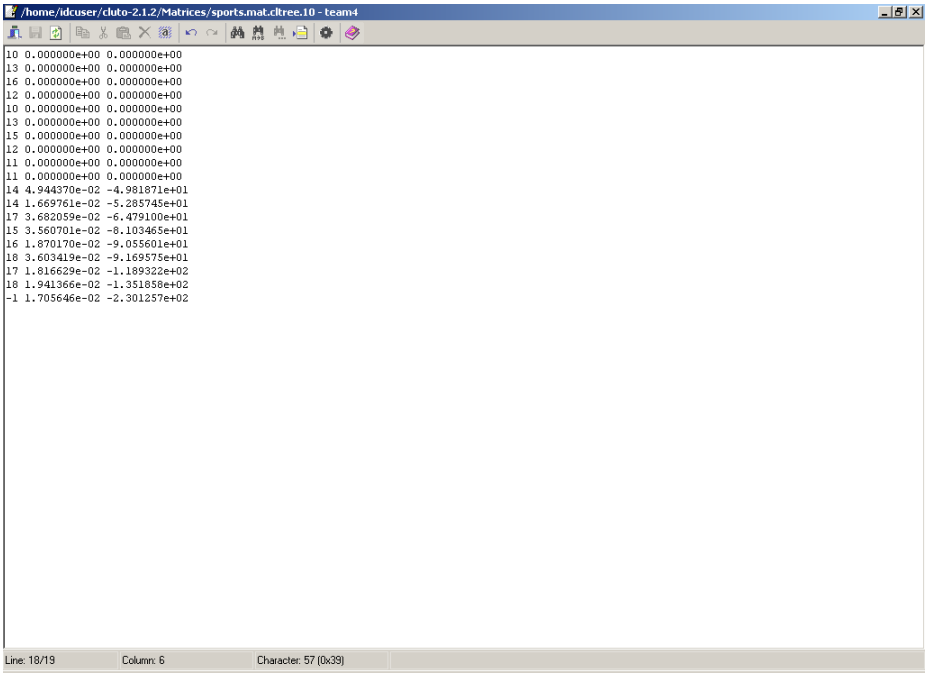


Figure 5: Tree file format

xiii. Other Features

Two additional features are provided by CLUTO:

a. gCLUTO

- gCLUTO is a cross-platform graphical application for clustering low- and high-dimensional datasets and for analyzing the characteristics of the various clusters as shown in figure 5.

- gCLUTO provides tools for visualizing the resulting clustering solutions using tree, matrix, and an OpenGL-based mountain visualization as shown in figure 4.

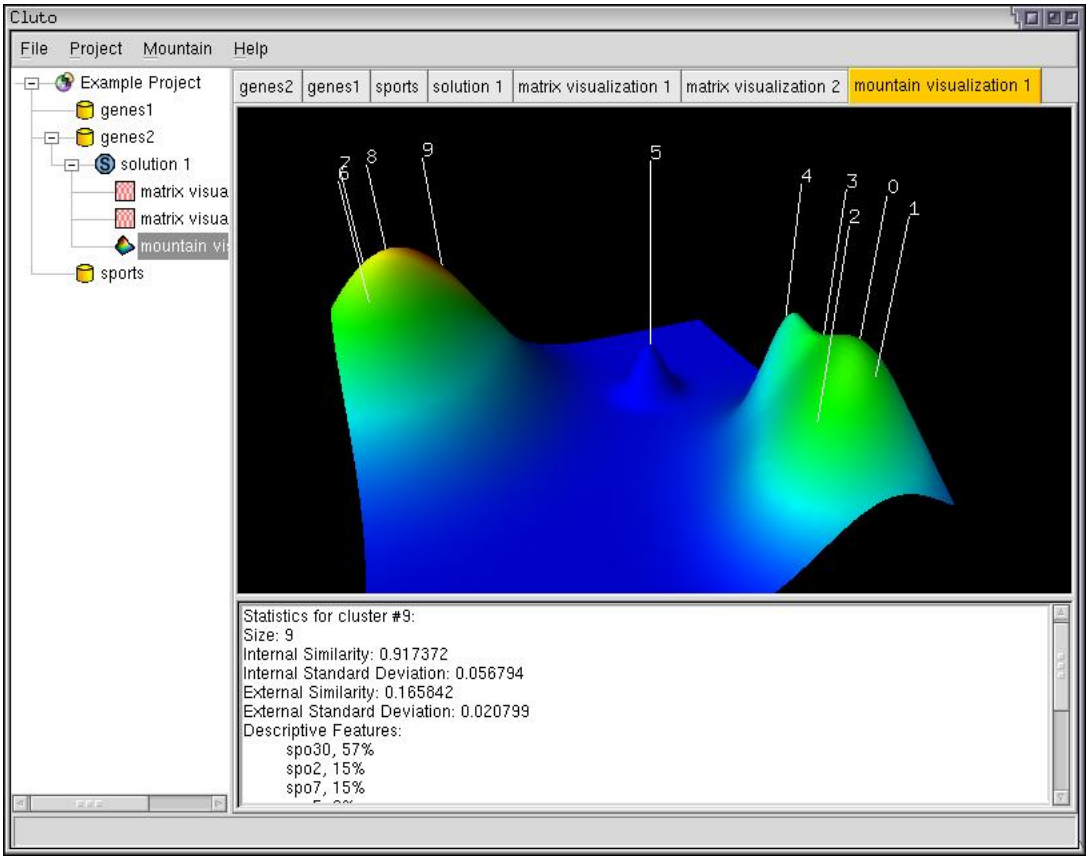


Figure 6: gCLUTO (Mountain View)

Adapted from: http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/overview

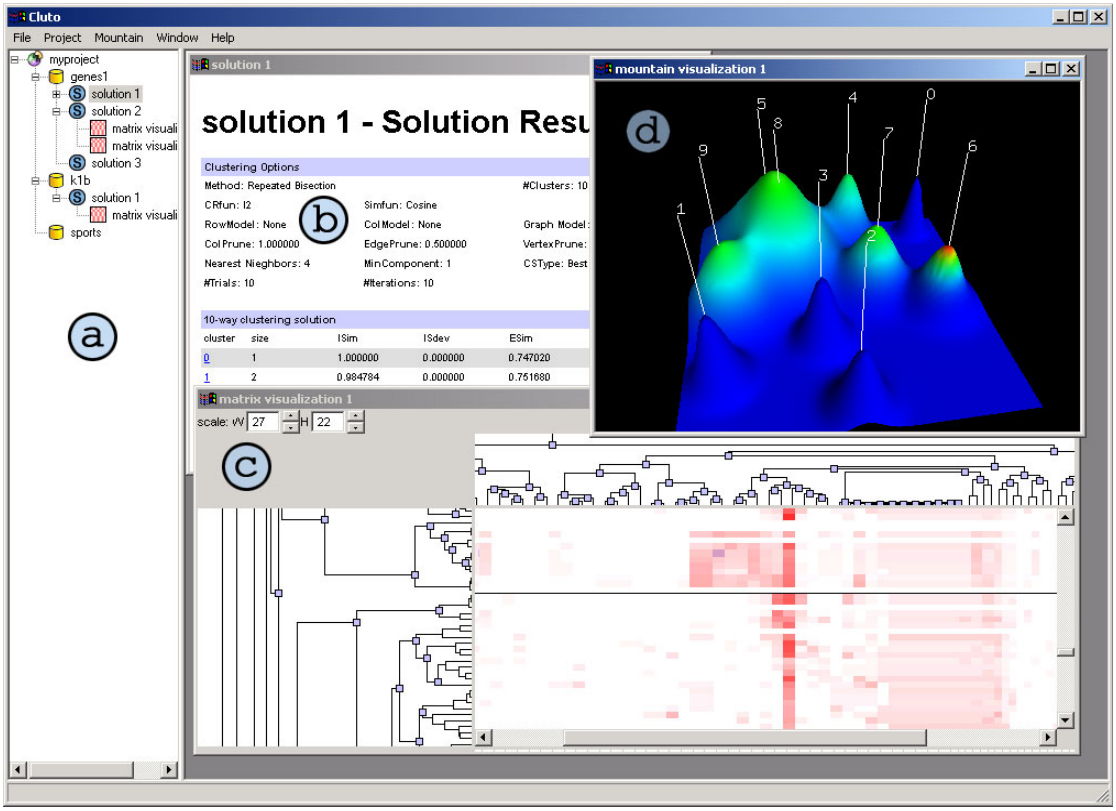


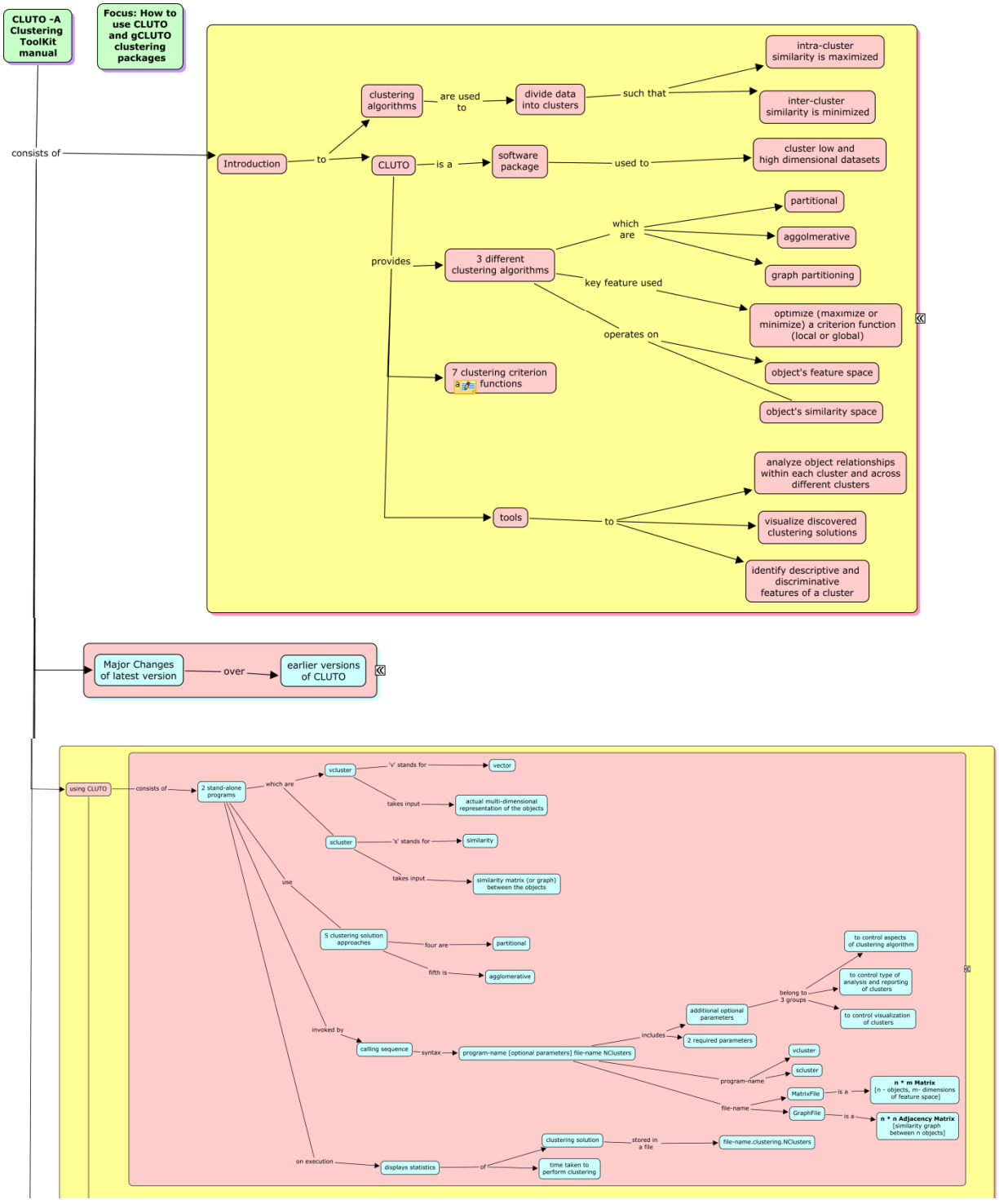
Figure 7: wCLUTO (Matrix Visualization)

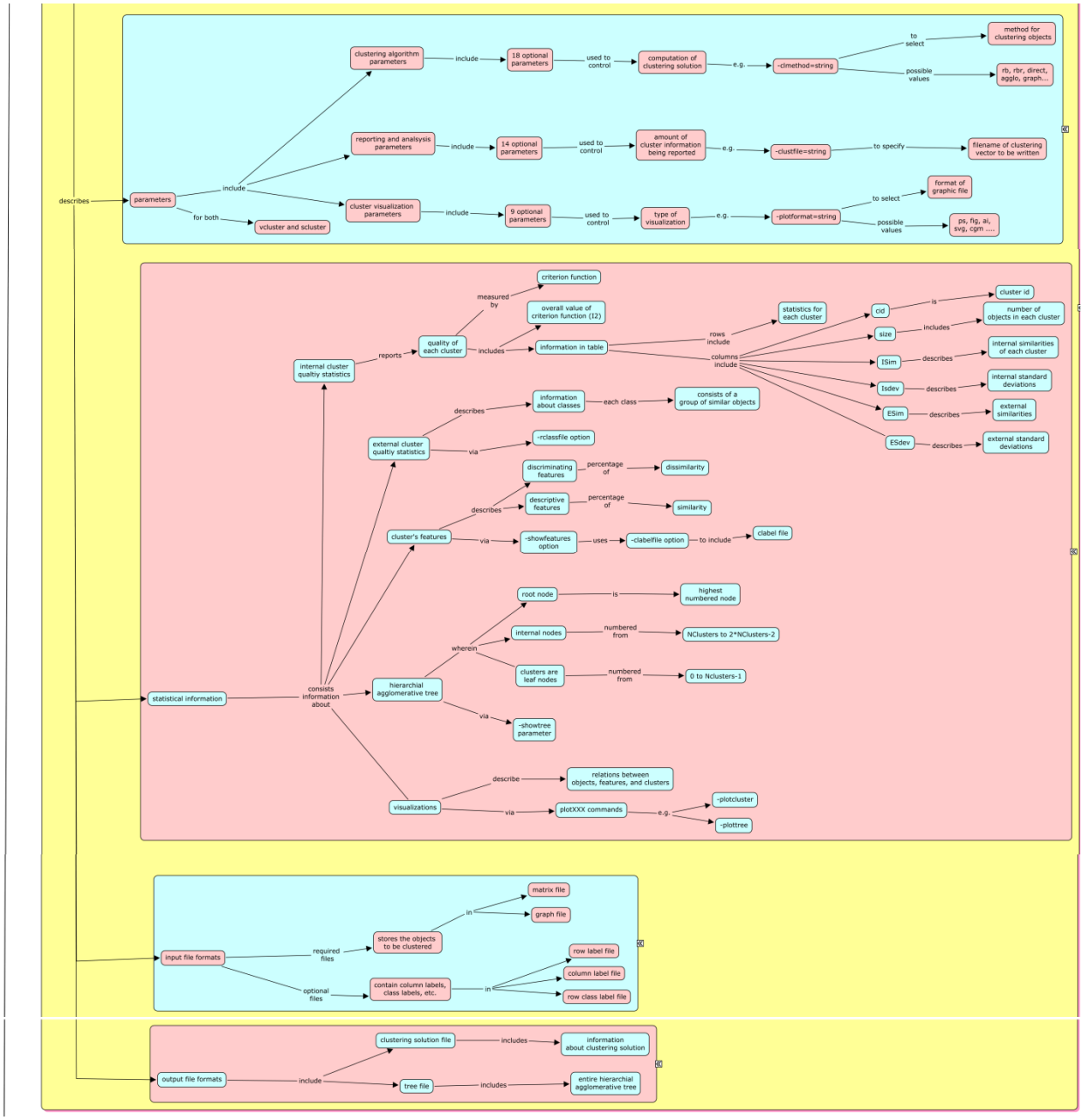
Adapted from: http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/overview

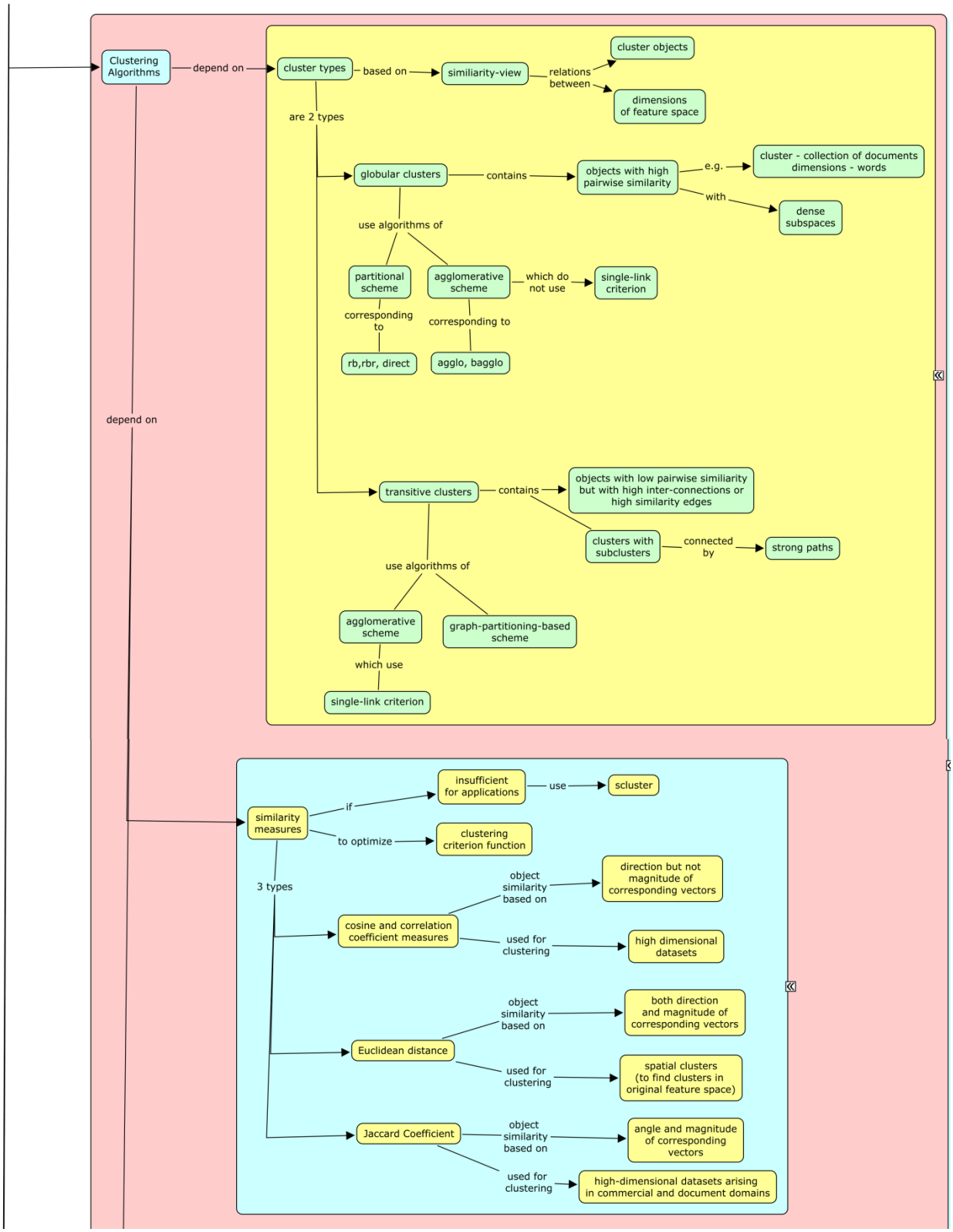
b. wCLUTO

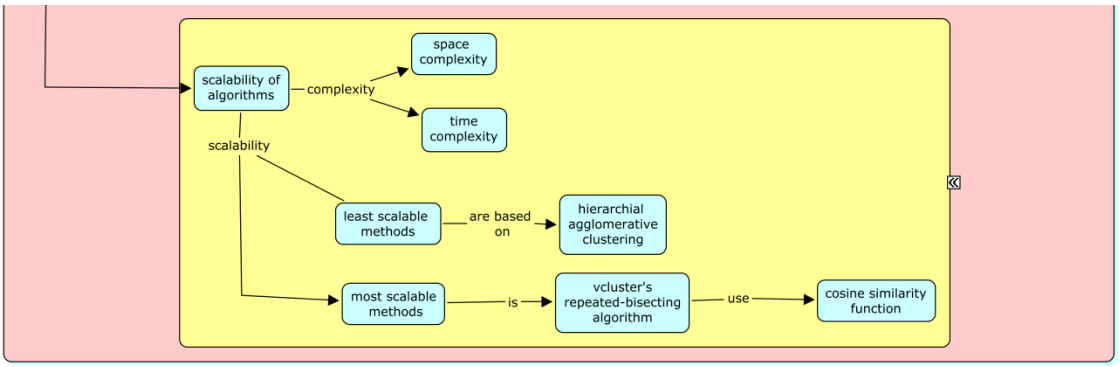
- wCLUTO is a web-enabled data clustering application that is designed for the clustering and data-analysis requirements of gene-expression analysis.
- Users can upload their datasets, select from a number of clustering methods, perform the analysis on the server, and visualize the final results.
- The wCLUTO web-server is hosted by the Center of Computational Genomics and Bioinformatics at the University of Minnesota.

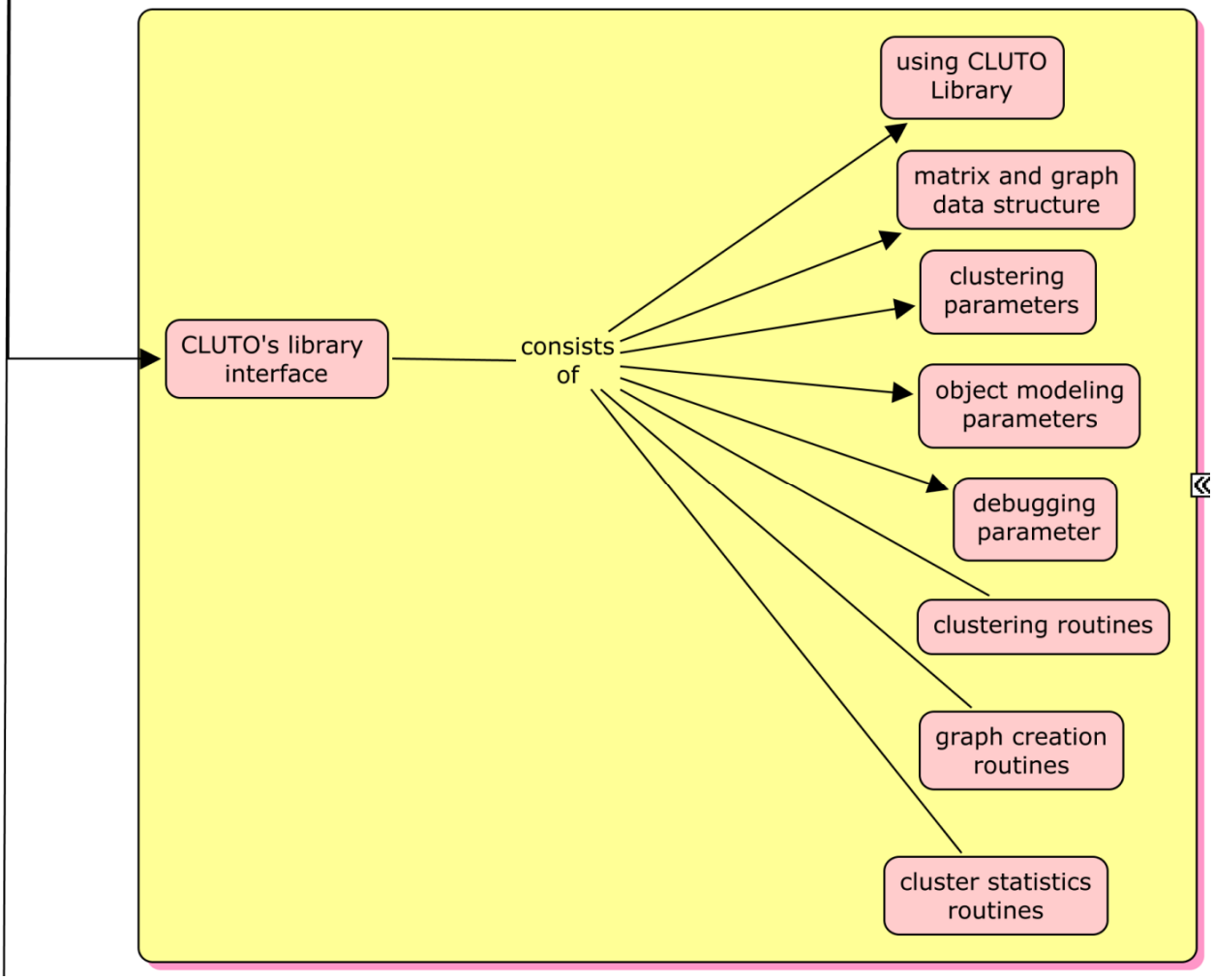
10) Concept Map

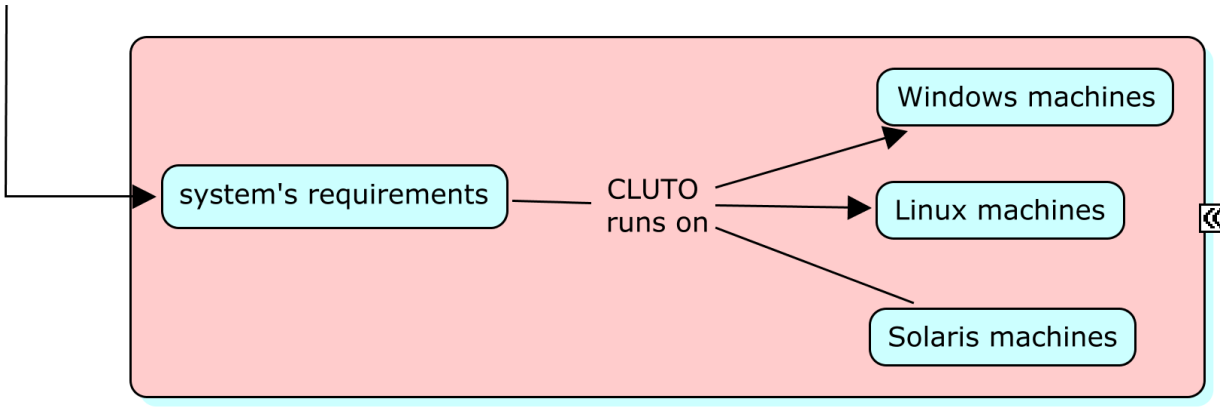



describes
parameters
include
for both
vcluster and scluster
clustering algorithm parameters
include
18 optional parameters
used to control
computation of clustering solution
e.g.
-clmethod=string
to select
method for clustering objects
possible values
rb, rbr, direct, agglo, graph...
reporting and analsysis parameters
include
14 optional parameters
used to control
amount of cluster information being reported
e.g.
-clustfile=string
to specify
filename of clustering vector to be written
cluster visualization parameters
include
9 optional parameters
used to control
type of visualization
e.g.
-plotformat=string
to select
format of graphic file
possible values
ps, fig, ai, svg, cgm ....
statistical information
consists information about
internal cluster qualtiy statistics
reports
quality of each cluster
measured by
criterion function
includes
overall value of criterion function (I2)
information in table
rows include
statistics for each cluster
columns include
cid
is
cluster id
includes
number of objects in each cluster
size
ISim
describes
internal similarities of each cluster
Isdev
describes
internal standard deviations
ESim
describes
external similarities
ESdev
describes
external standard deviations
external cluster qualtiy statistics
describes
information about classes
each class
consists of a group of similar objects
via
-rclassfile option
cluster's features
describes
discriminating features
percentage of
dissimilarity
descriptive features
percentage of
similarity
via
-showfeatures option
uses
-clabelfile option
to include
clabel file
hierarchial agglomerative tree
wherein
root node
is
highest numbered node
internal nodes
numbered from
NClusters to 2*NClusters-2
clusters are leaf nodes
numbered from
0 to Nclusters-1
via
-showtree parameter
visualizations
describe
relations between objects, features, and clusters
via
plotXXX commands
e.g.
-plotcluster
-plottree
input file formats
required files
stores the objects to be clustered
in
matrix file
graph file
optional files
contain column labels, class labels, etc.
in
row label file
column label file
row class label file
output file formats
include
clustering solution file
includes
information about clustering solution
tree file
includes
entire hierarchial agglomerative tree


Clustering Algorithms
depend on
cluster types
based on
similiarity-view
relations between
cluster objects
dimensions of feature space
are 2 types
globular clusters
contains
objects with high pairwise similarity
e.g.
cluster - collection of documents
dimensions - words
with
dense subspaces
use algorithms of
partitional scheme
agglomerative scheme
which do not use
single-link criterion
corresponding to
corresponding to
rb,rbr, direct
agglo, bagglo
depend on
transitive clusters
contains
objects with low pairwise similiarity but with high inter-connections or high similarity edges
clusters with subclusters
connected by
strong paths
use algorithms of
agglomerative scheme
graph-partitioning-based scheme
which use
single-link criterion
similarity measures
if
insufficient for applications
use
scluster
to optimize
clustering criterion function
3 types
object similarity based on
direction but not magnitude of corresponding vectors
cosine and correlation coefficient measures
used for clustering
high dimensional datasets
object similarity based on
both direction and magnitude of corresponding vectors
Euclidean distance
used for clustering
spatial clusters (to find clusters in original feature space)
Jaccard Coefficient
object similarity based on
angle and magnitude of corresponding vectors
used for clustering
high-dimensional datasets arising in commercial and document domains


scalability of algorithms
complexity
space complexity
time complexity
scalability
least scalable methods
are based on
hierarchial agglomerative clustering
most scalable methods
is
vcluster's repeated-bisecting algorithm
use
cosine similarity function


CLUTO's library interface
consists of
using CLUTO Library
matrix and graph data structure
clustering parameters
object modeling parameters
debugging parameter
clustering routines
graph creation routines
cluster statistics routines



11) Exercises/Learning Activities

i. **In-Class exercise 1** (10 minutes)

a. This exercise will help you understand the basic concepts of clustering.

The links below contain demos to sample clustering algorithms:

- Sample clustering algorithm using K-means : http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html
- Sample hierarchical clustering algorithm: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

b. Select 100 for data size and 5 for the number of clusters and then click on the **Initialize** button to generate them in random positions.

c. Click on **Run** to begin the simulation. During simulation data positions are fixed.

d. Save the results of each experiment as image and add descriptive comments.

ii. **In-Class exercise 2** (30 minutes)

a. Go to the website of the Center for Machine Learning and Intelligent Systems http://archive.ics.uci.edu/ml/datasets/Iris

b. Download the IRIS dataset.

c. Convert the IRIS dataset format to CLUTO format (.mat file) either manually or using program doc2mat.

d. Use the vcluster program to cluster the input dataset into three clusters.

e. Build and display a hierarchical agglomerative tree.

f. Calculate and display the discriminative and descriptive features of each cluster.

g. Every iris has four features (sepal length in cm, sepal width in cm, petal length in cm, petal width in cm).

iii. **In-Class Exercise 3** (15 minutes)

Use the sports.mat file provided in the cloud to answer the following questions.

a. Cluster the input dataset into 10 clusters using the vcluster program and cosine similarity function.

b. What clustering algorithm does vcluster use by default?

c. Use a partitional and an agglomerative clustering algorithm to perform the clustering.

d. Compute clustering by combining both the partitional and agglomerative methods used in c.

e. Use the given class file sports.rclass to compute the quality of clustering solution used (Use algorithms used in c. and d.).

f. Analyze the output. Which internal characteristic displays internal similarities and which internal characteristic displays external similarities?

12) Evaluation of Learning Outcomes

i. Are students familiar with the fundamental concepts, applications and techniques of Clustering?

ii. Are students capable of running clustering programs using CLUTO toolkit? This can be determined if a student can successfully run a vcluster/scluster program using appropriate parameters over a given input file and produce an output file with a desired number of clusters.

13) Glossary

i. **Cluster**: A group of objects grouped together by a criterion (similarity) function, often a set of documents.

ii. **Feature:** Information extracted from an object and used during query processing.

iii. **Index:** A data structure built for the images to speed up searching.

iv. **Information Retrieval:** Field of computer science which studies the retrieval of information (not data) from a collection of written documents.

v. **Metadata:** Informative data about documents or input data. Metadata includes attributes of documents such as author name, date of creation, year of publication, etc.

vi. **Repository:** A physical or digital place where objects are stored for a period of time, from which individual objects can be obtained if they are requested.

vii. **Vcluster:** Stand-alone program that takes as input the actual multi-dimensional representation of objects to be clustered.

viii. **Scluster:** Stand-alone program that takes as input the similarity matrix (or graph) between the objects to be clustered.

ix. **Criterion function**: A function that maps an event onto a real number representing the economic cost or regret associated with the event.

x. **Similarity function**: A function that defines what objects would belong to a particular cluster.

xi. **Feature vector:** An n-dimensional vector of numerical features representing an object.

xii. **Feature space:** The vector space associated with feature vectors.

xiii. **Euclidean Distance:** The ordinary distance between two points in a vector space.

xiv. **Partitional algorithms:** Begin with the whole set and proceed to divide it into successively smaller clusters.

xv. **Agglomerative algorithms:** Begin with each element as a separate cluster and merge them into successively larger clusters.

14) References

i. Weblinks:

[1] Home Page: http://glaros.dtc.umn.edu/gkhome/views/cluto

[2] Manual: http://glaros.dtc.umn.edu/gkhome/fetch/sw/cluto/manual.pdf

[3] Download Page: http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download

[4] Publications: http://glaros.dtc.umn.edu/gkhome/cluto/cluto/publications

ii. Related research publications:

[5] A Segment-based Approach To Clustering Multi-Topic Documents. *Andrea Tagarelli and George Karypis. Text Mining Workshop, SIAM Datamining Conference, 2008.* Department of Computer Science and Engineering, University of Minnesota, Minneapolis. TR #08-004.

[6] Hierarchical Clustering Algorithms for Document Datasets. Ying Zhao, Goerge Karypis, Usama Fayyad. Data Mining and Knowledge Discovery. Boston:Mar 2005. Vol. 10, Iss. 2, p. 141-168, Department of Computer Science, University of Minnesota, Minneapolis. TR #03-027.

[7] Topic-Driven Clustering for Document Datasets. *Ying Zhao and George Karypis. SIAM International Conference on Data Mining, pp. 358-369,2005.* Department of Computer Science, University of Minnesota, Minneapolis. TR #05-017.

[8] Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering. *Ying Zhao and George Karypis. Machine Learning, 55, pp. 311-331, 2004.*

[9] Clustering in Life Sciences. *Ying Zhao and George Karypis. In Functional Genomics: Methods and Protocols, M. Brownstein, A. Khodursky and D. Conniffe (editors). Humana Press, 2003.*

[10] Evaluation of Hierarchical Clustering Algorithms for Document Datasets. *Ying Zhao and George Karypis. 11th Conference of Information and Knowledge Management (CIKM), pp. 515-524, 2002.*

[11] Criterion Functions for Document Clustering: Experiments and Analysis. *Ying Zhao and George Karypis. Department of Computer Science, University of Minnesota, Minneapolis, MN, 2001. TR #01-40.*

[12] A Comparison of Document Clustering Techniques. *Michael Steinbach, George Karypis and Vipin Kumar. KDD Workshop on Text Mining, 2000. Department of Computer Science / Army HPC Research Center, University of Minnesota, Minneapolis. TR #00-034.*

[13] CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling. *George Karypis, Eui-Hong Han, Vipin Kumar. IEEE Computer 32(8): 68-75, 1999.*

iii. Additional reading list for students:

[14] CURE: An efficient clustering algorithm for large databases. *Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. In Proceedings of 1998 ACM-SIGMOD International Conference on Management of Data, 1998. Page: 73. Place: New York, USA.*

[15] hMETIS 1.5: A hypergraph partitioning package. *G. Karypis and V. Kumar. Technical report, Department of Computer Science & Engineering, Army HPC Research Center, University of Minnesota, Minneapolis, 1998.*

[16] METIS 4.0: Unstructured graph partitioning and sparse matrix ordering system. *G. Karypis and V. Kumar. Technical report, Department of Computer Science, University of Minnesota, Minneapolis, 1998.*

15) Contributors:

Authors:

CS_5604: Information Storage and Retrieval- Team 4:

i. Vijay, Sony

ii. El Meligy Abdelhamid, Sherif

iii. Malayattil, Sarosh

Reviewers:

Dr. Edward Fox