Principles of Knowledge Discovery in Data

Fall 2000

Assignment 1

5%

Systems List

1. **Alice** (Isoft - http://www.isoft.Fr)
2. **Clementine** (SPSS - http://www.spss.com/software/clementine/)
3. **DBDiscover** (University of Regina - H. Hamilton - http://www.cs.uregina.ca)
4. **DBMiner** (Simon Fraser University - J. Han - http://db.cs.sfu.ca)
5. **PolyAnalyst** (Megaputerr - http://www.megaputer.com)
6. **Intelligent miner** (formerly known as quest) (IBM - http://www.software.ibm.com/data/iminer/)
7. **MineSet** (SGI - http://www.sgi.com/software/mineset/)
8. **Easyminer** (MINEit - http://www.mineit.com)
9. Discipulus
10. **SAS** (SAS institute - http://www.sas.com)
11. UCI database (KDD Archives)

Assignment 1

6 groups

Deliverable:

- Write a report about the data mining/data warehousing system you selected. Explain the functionalities, capabilities, and limitations. Classify the system according to the classification scheme we discussed in class.
- Convert this report into HTML so that we can publish it on the course web site.
- Present your findings in a 15 minute class talk.

Due Dates:

Report due date: October 20th 2000
Presentation dates: October 23th and 25th 2000.

Responsibilities

It is the students responsibility to form groups.

Task distribution inside group is the group's responsibility.

Suggestion: parallel + merge in report and presentation

Presentation: one student per group.

Evaluation per group.

Principles of Knowledge Discovery in Databases

Fall 2000

Projects

35%

Implementation vs. Survey

Implementation

Each student should attempt to implement one project
- Improvements on existing algorithms
- Combination of existing algorithms
- New scalable algorithms for mining, data cube implementations, indexing, dimension reduction, etc.
- Integration of knowledge discovery and statistic techniques

Survey

Survey papers should summarize previous research and report on recent research issues and advances in the chosen topic.
20 to 30 pages.

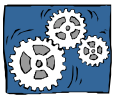
Due Dates:
Proposal: October 11th, 2000
Report: December 4th, 2000
Presentation + Demo: November 27th – December 8th 2000

Survey Topic Examples

1. State of the art in speech data mining
2. State of the art in image data mining
3. State of the art in video data mining
4. State of the art in parallel data mining

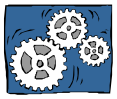
Project Examples

1. Writing educational Java applets to illustrate data mining and data warehousing concepts (classification trees, clustering, etc.).
2. Implementing a visualization module to visualize datacube data (or data mining results) on the projection screens of the VizRoom (in 3D) .
3. Implementing an incremental, a distributed or parallel mining algorithm (many possibilities) .
4. Implementing a similarity search in time series data.
5. Implementing OLAM algorithms (using data cubes as input) (many possibilities) .
6. Implementing multi-level association rule mining.
7. Implementing classification algorithms (many possibilities).



Project Examples (con't)

8. Implementing clustering algorithms (many possibilities).
9. Implementing discretization of numerical attributes and concept hierarchy building.
10. Extracting interesting metadata about web documents.
11. Finding authoritative web documents (such as CLEVER).
12. Implementing clustering of Web documents according to document attributes, usage and content.
13. Implementing transaction and sequence identification in web access logs.
14. Implementing association rule extraction from web access logs.
15. Implementing web document restructuring.



Project Examples (con't)

16. Implementing Web document browsing à la OLAP using existing ontologies.
17. Designing a query language for querying interesting rules.
18. Implementing clustering of access patterns in web access logs.
19. Text mining from text documents (e-mails, memos, on-line news, etc.).
20. Implementing association rule extraction from images.
21. Implementing visualization techniques for displaying data mining results or for helping interactive data mining.

Project Examples (con't)

22. Extract key information from text (summarization, keywording).
23. Organize documents by subjects (classification).
24. Find predominant themes in a collection of documents (clustering).
25. Clustering images based on feature localization.

Proposal by October 11th

1 to 2 pages

– Survey ➔ topic, initial references, schedule.

– Implementation ➔ project topic, implementation choices, approach and schedule.

Implementations could be with C/C++ or Java, on Linux, Window NT/98 , or other systems.