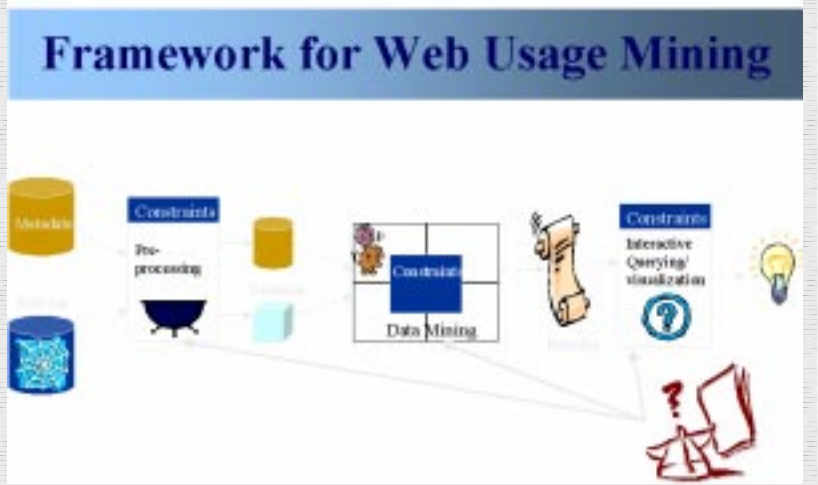
CMPUT 695 course project
Web Site Miner

Andrew Foss, Weinan Wang

Big picture of Web Usage Mining



Web log data preprocessing

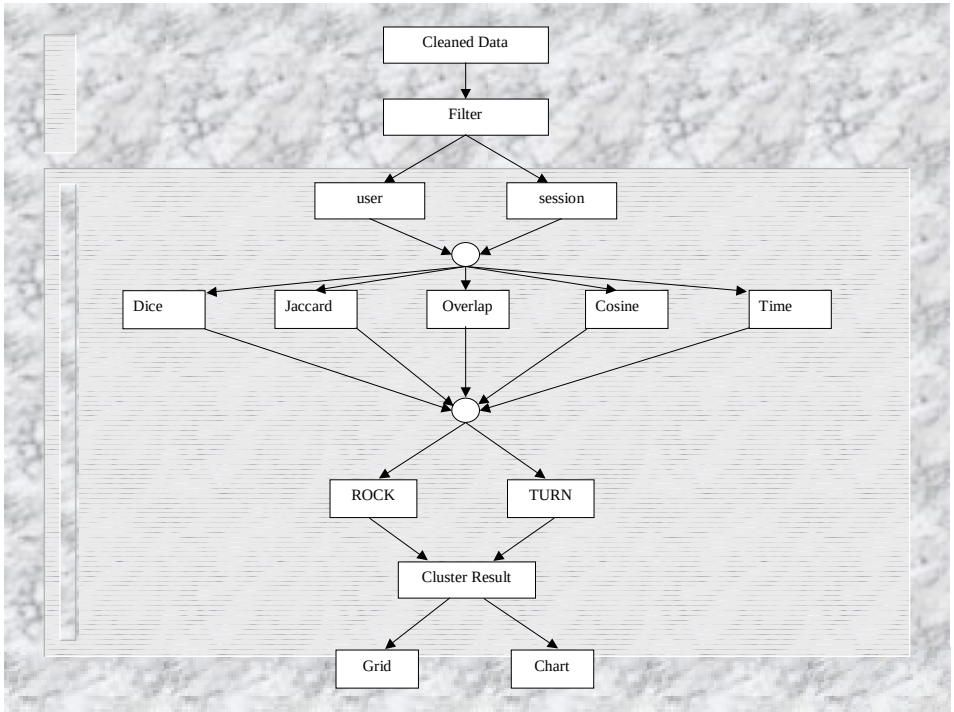
- Clean the web log.
- Identify user sessions.
- Create ids for web pages accessed

Session	page#	time stamp
01000102	962	945458058
01000102	962	945458060
01000102	483	945458060
01000102	484	945458060
01100001	965	937344265
01100001	963	937340669
01100001	964	937340670
01100001	964	937341439

Page#	page ID	URL
6 :	0050	: /Courses/TECH142/RequiredResources/index.html
7 :	0060	: /Courses/TECH142/TeachingStaff/index.html
8 :	007	: /Courses/TECH142/index.html
9 :	008	: /Courses/TECH142/side.html
10 :	01	: /Courses/TECH150
11 :	0100	: /Courses/TECH150/CourseDescription/index.html
12 :	0110	: /Courses/TECH150/Evaluation/index.html
13 :	0120	: /Courses/TECH150/Expectations/index.html
14 :	0130	: /Courses/TECH150/LearningSupport/index.html

What we did in this project

- Filtering and collecting session information;
- Clustering over session and user (ROCK and TURN);
- User interface development.



Filtering and collecting session information

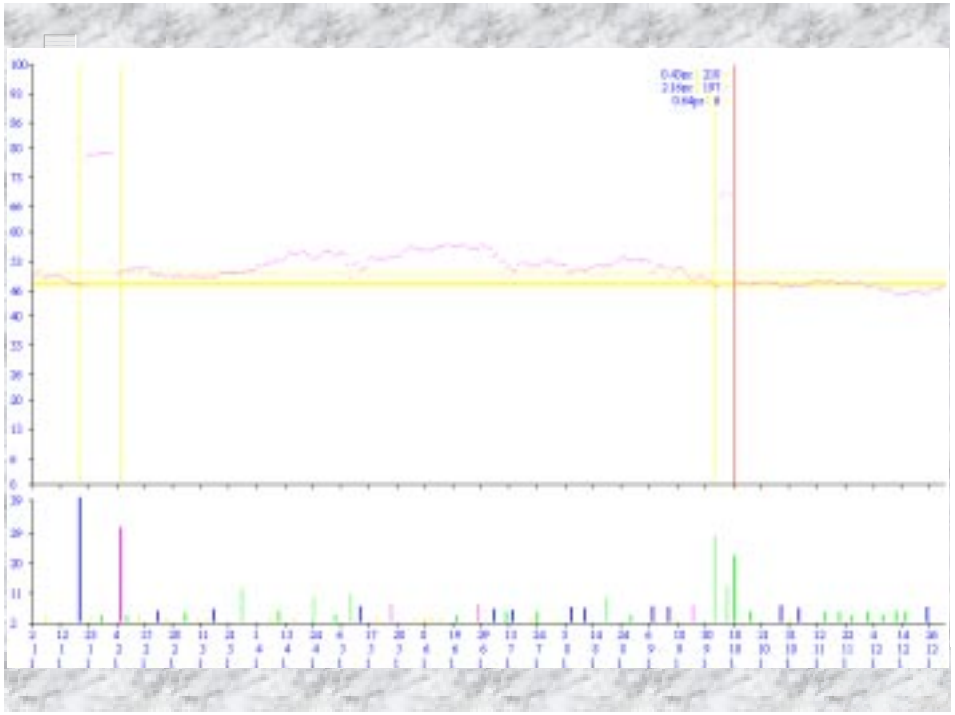
- Remove duplicate pages
- Remove duplicates to levels
- Stems only
- Maximal Forward References
- Remove short transactions (user definable)

Viewing the data

- View individual session/user
- Compare individual sessions, see distance, results of filters
- See other sessions in comparison to any individual session

TURN

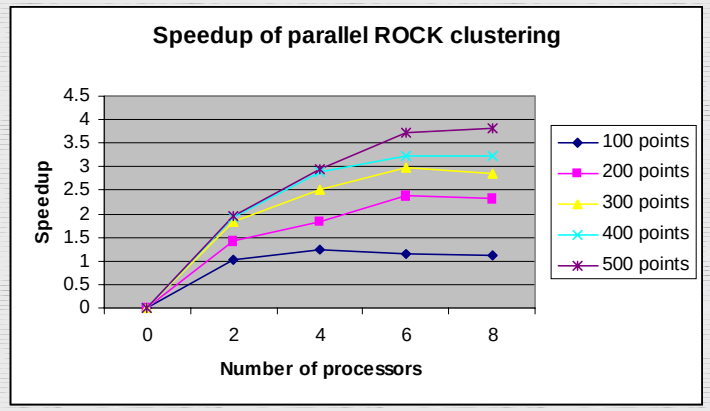
- Discovering parameters for ROCK
- Clustering
- Looks at turning points in the distances
- Found using third differential in the series
- Application to Euclidean data like Time Series
- A work in progress



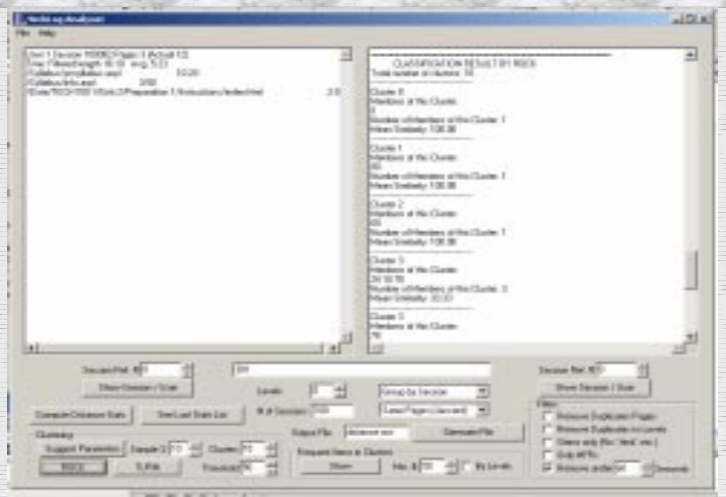
Session and user clustering by ROCK

- Measure similarity of sessions and users -- Jaccard coefficient, Dice coefficient, Overlap coefficient, Cosine coefficient, Time.
- User specified sample rate.
- Collect outliers in classification part.
- Parallel ROCK

Speedup of parallel ROCK



Our product -- WebSiteMiner



Mining results visualization

- Finding, filtering and sorting frequent items
- Spread sheet style view of frequent items in clusters
- Chart view of frequent items in clusters

Further work

- Open any web log and clean data.
- More mining functionality -- finding association rules using neural networks, Bayesian network.
- Printing, user manual, etc.