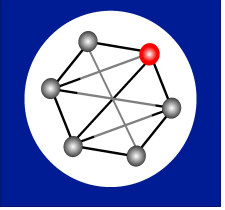


DBTechNet

DBTech Pro Workshop

Knowledge Discovery from Databases (KDD)
Including Data Warehousing and Data Mining

Dimitris A. Dervos
dad@it.teithe.gr
http://aetos.it.teithe.gr/~dad

Georgios Evangelidis
gevan@uom.gr
http://users.uom.gr/~gevan/

Friedrich Laux
Friedrich.Laux@Reutlingen-University.de

Helsinki

January 2009

Last updated: Jan 09, 2008

Part 1

An Introduction to

Data Warehousing & OLAP

DW & OLAP Tutorial

- Existing tutorial covers ETL Process, Data Warehousing issues (architectures and implementations), OLAP operators, advanced issues like indexing, etc.
- We will update it and include latest developments, like OLAP with web-based GUI

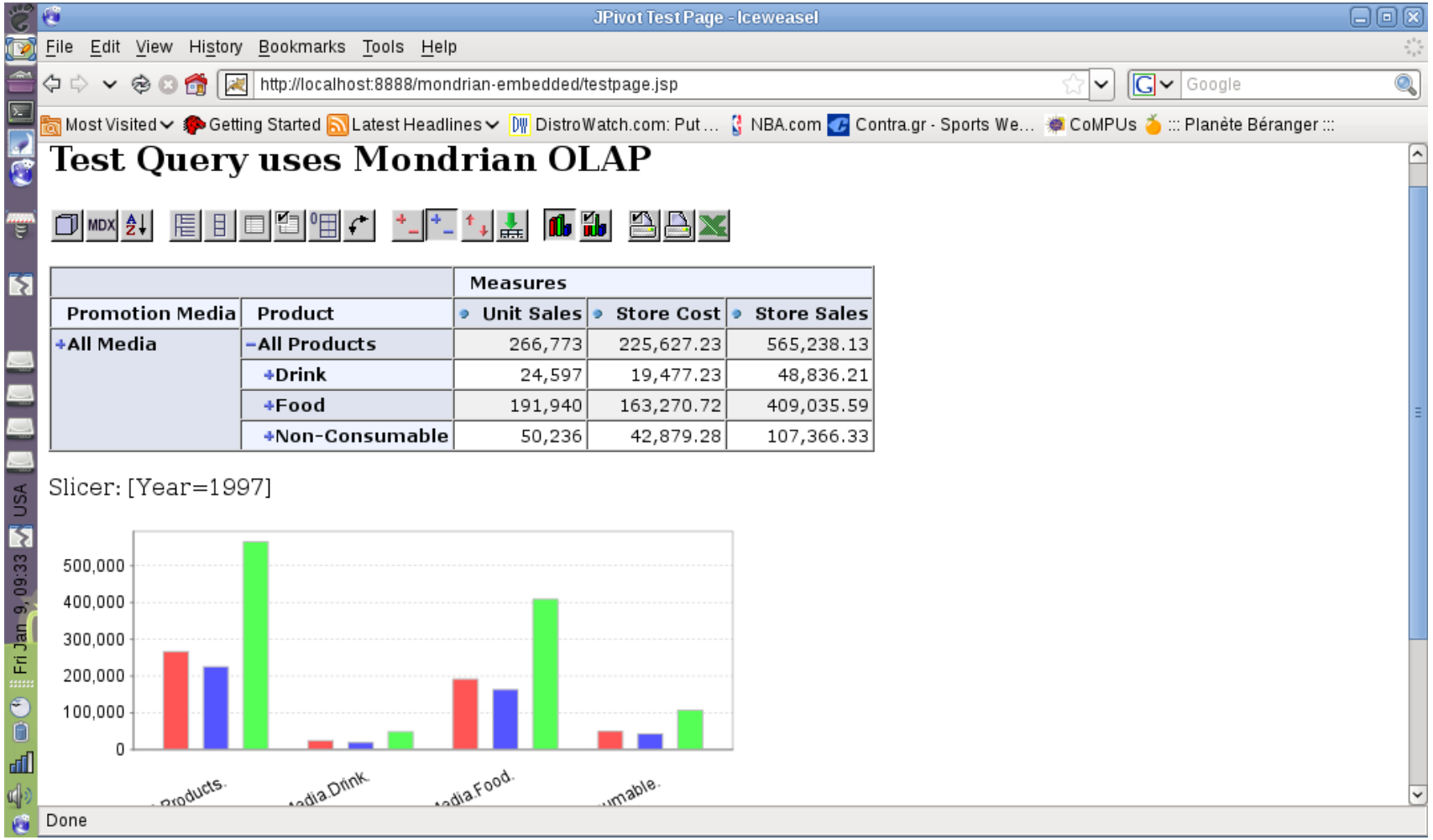
DW & OLAP Tutorial: Self-Study Lab Material

- Existing self-study guide and exercise from DBTech Pro Project
- It uses MS SQL Server for ETL & MS Analysis Services for DW construction and OLAP
- We will update it to use new version of the software (SQL Server 2005)

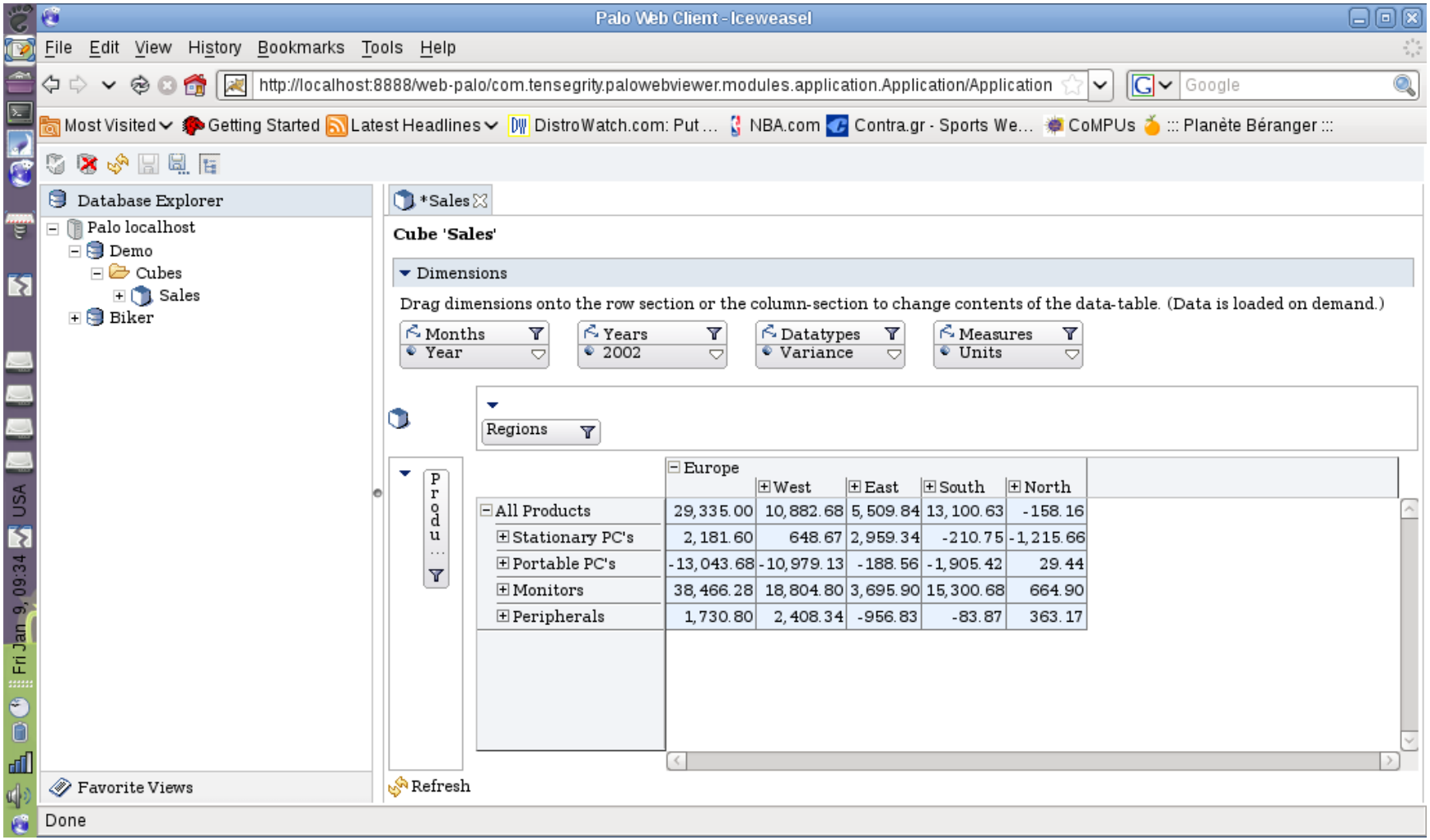
DW & OLAP Tutorial: New Material

- Self-study material and exercises for OSS solutions (one or both of PALO and Mondrian Pentaho)
- Both use web-based gui clients to operate on the OLAP cubes (J2EE architecture)
- PALO is MOLAP
- Mondrian Pentaho is ROLAP (MySQL)

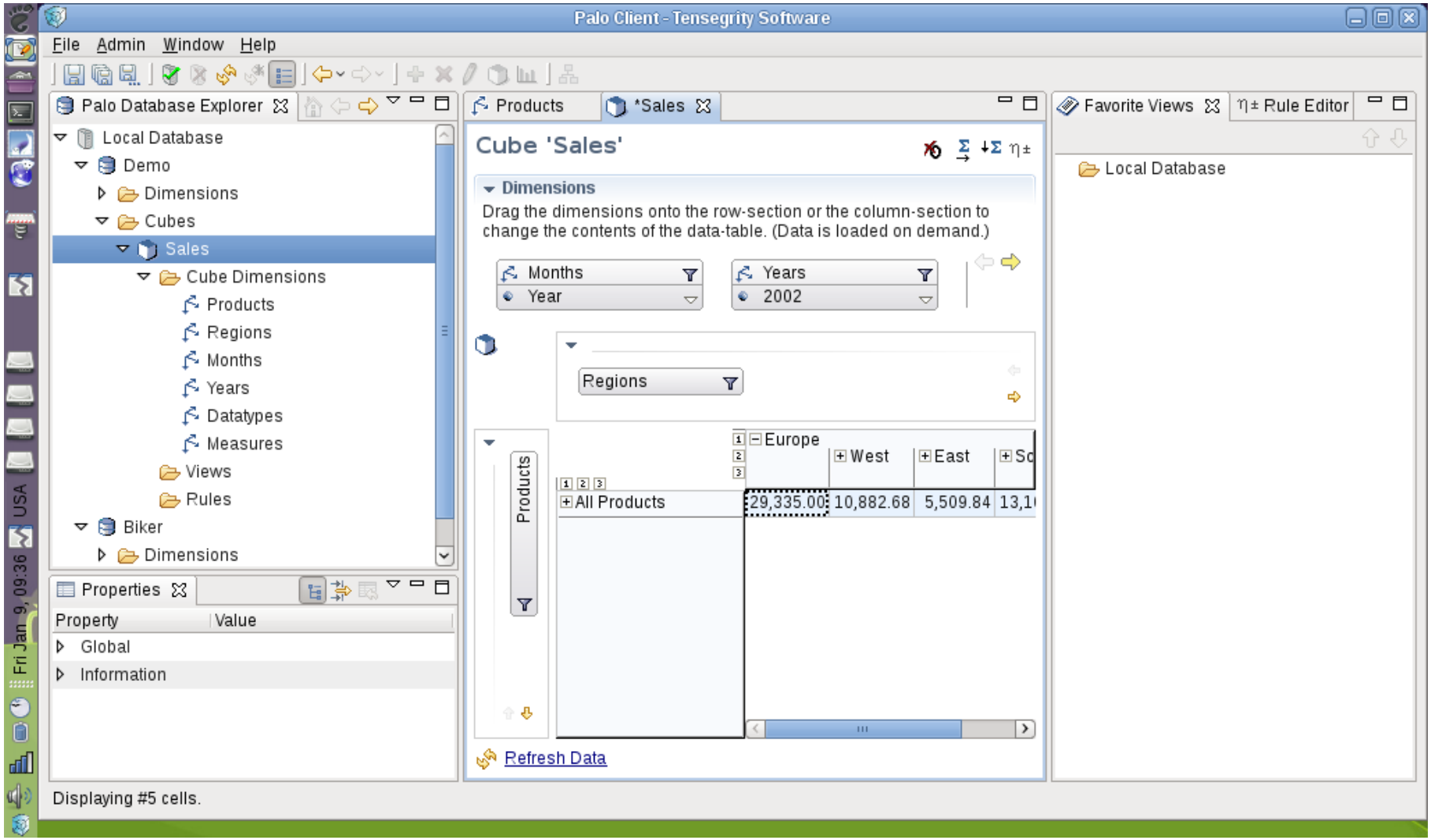
DW & OLAP Tutorial: Mondrian Pentaho web client



DW & OLAP Tutorial: PALO Web Client



DW & OLAP Tutorial: Palo Java Client



Part 2

An Introduction to

Knowledge Discovery from Databases & Data Mining

Tutorial

- Data – Information – Knowledge
- DM queries vs. SQL
- Data Mining Strategies
- Information as Entropy
- Supervised Learning / Classification: Decision Trees
- Unsupervised Clustering
- Affinity Analysis / Association Rules
- Bringing it together (on information representation)
  - rules with exceptions
  - attribute-to-attribute relations
  - rules that imply other rules
  - association vs. classification rules
  - probabilistic clustering
  - dendrograms

Hands-on (Virtual) Lab Content

Case study & Self-study exercises with model answers

- Mining your Business in Retail with IBM DB2 Intelligent Miner
- K-Means Clustering with WEKA
- Classification with the WEKA User Classifier

Laboratory exercises

- Association rules mining with IBM DB2 Intelligent Miner
- Classification (Decision Tree) with WEKA

Mining your Business in Retail (IBM tutorial)

Utilizes SQL and Easy Mining Procedures for the IBM Intelligent Miner®
Retrieved from: **http://www.ibm.com/developerworks/edu/dm-dw-dm-retail_tutorial-i.html**

Association Rules Mining

- Data exploration phase
- Data preparation phase (SQL views)
- Association rules model building
- Evaluation phase: IM Visualization
- Evaluation phase: the rules (SQL) view
- Deployment phase: products recommendation

Clustering

- Data preparation phase (SQL views)
- Model building-1: Demographic Clustering
- Evaluation phase: IM Visualization
- Evaluation phase: results interpretation
- Model building-2: an improved clustering model
- Evaluation phase-2: results interpretation, cluster analysis
- Deployment phase: improved product recommendations

K-Means Clustering with WEKA*

- Retrieved from **http://maya.cs.depaul.edu/~classes/ect584/WEKA/index.html**
- Bank dataset
- Result output with 'Cluster' attribute exported for further processing (e.g. classification mining)
- May evaluate the performance of K-Means for different input parameters (seeds)

---

* WEKA is open source data mining software developed by the University of Waikato, New Zealand (**http://www.cs.waikato.ac.nz/ml/weka/**)

Classification with the WEKA User (tree) Classifier

- WEKA supplied visual image data set segmented into classes such as grass, sky, foliage, brick, and cement based on attributes giving average intensity, hue, size, position, and various simple textural features.
- Classification proceeds manually up to a certain stage, utilizing any one of the available tree classification algorithms thereafter

Laboratory Exercise: Association rules mining

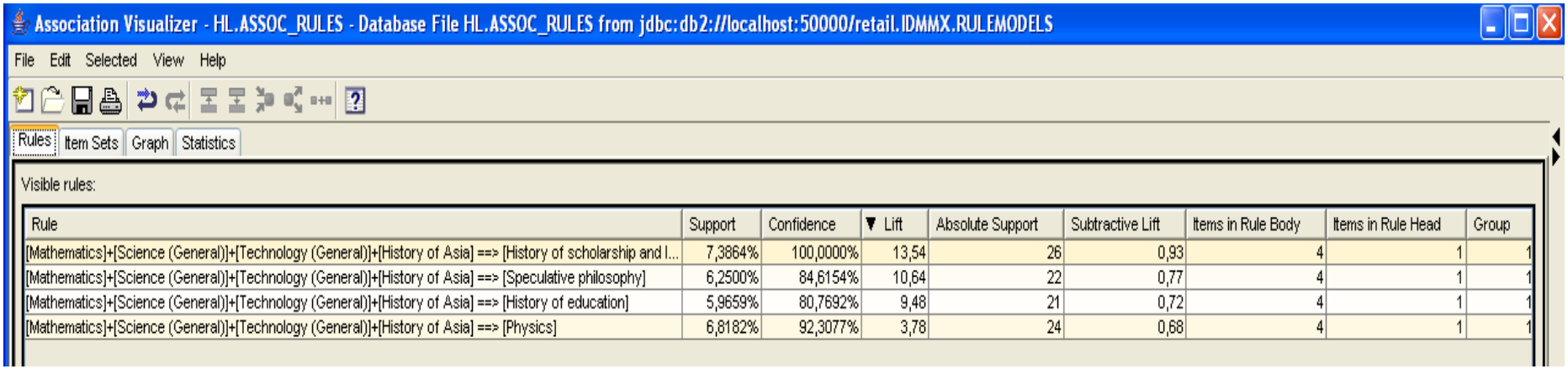
HEALlink (*http://www.heal-link.gr/journals/en/*)

HEALlink: Journal_Stats snapshot

HEALLINK - DB2 - HEALINK2 - DB2ADMIN.JOURNAL_STATS

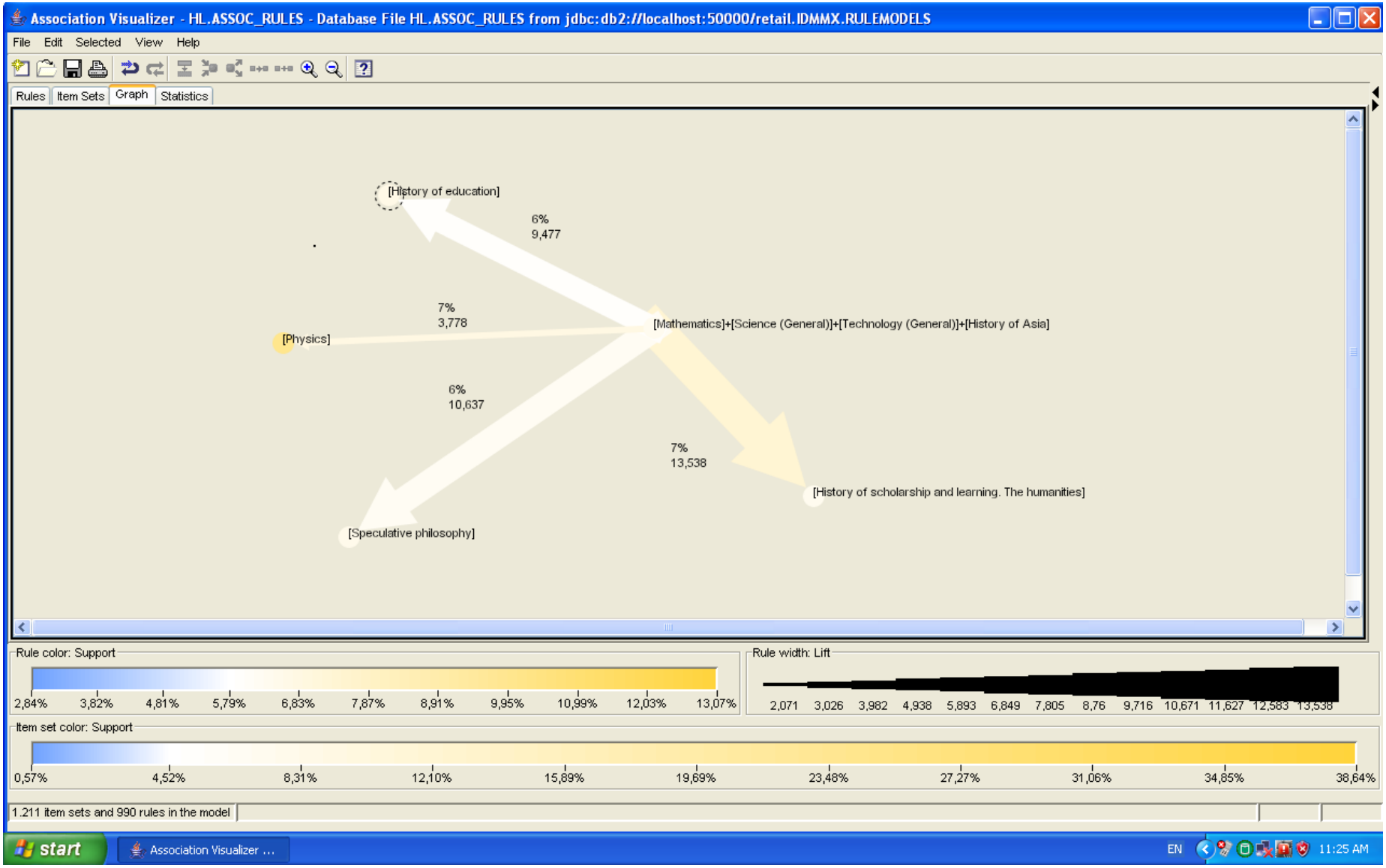
EMAIL	IP	TIME	TYPE	JOURNAL
lpispir@...	193.92.234...	Wed Mar 19 17:11:29 2003	subject	Computer Speech
lpispir@...	193.92.234...	Wed Mar 19 17:11:53 2003	subject	Lecture Notes in Computer Science
lpispir@...	193.92.234...	Wed Mar 19 17:12:25 2003	subject	IBM Systems Journal
lpispir@...	193.92.234...	Wed Mar 19 17:14:48 2003	subject	Wireless Personal Communications
lpispir@...	193.92.234...	Wed Mar 19 17:17:41 2003	subject	Photonic Network Communications
lpispir@...	193.92.234...	Wed Mar 19 17:18:47 2003	selected	Computer Communications
lpispir@...	193.92.234...	Wed Mar 19 17:20:34 2003	selected	Wireless Personal Communications
	193.92.234...	Wed Mar 19 17:25:26 2003		Computers in Biology and Medicine
	193.92.234...	Wed Mar 19 17:26:28 2003		Industrial Management
	193.92.234...	Wed Mar 19 17:26:57 2003		Behaviour
lpispir@...	193.92.234...	Sat Mar 22 13:39:28 2003	alphabetic	Wireless Personal Communications
lpispir@...	193.92.234...	Sat Mar 22 13:43:03 2003	selected	Computer Networks
	193.92.234...	Sun Mar 23 16:50:13 2003		Computers
	193.92.233...	Mon Mar 31 10:28:17 2003		Learning Environments Research
dad@it.t...	193.92.233...	Mon Mar 31 10:39:39 2003	selected	Information Processing
	193.92.234...	Mon Mar 31 20:09:33 2003		Journal of Educational Thought
	193.92.234...	Mon Mar 31 20:12:07 2003		Journal of Food Composition and Analysis
	193.92.233...	Tue Apr 1 7:40:02 2003		Journal of Biotechnology
joankw...	193.92.233...	Tue Apr 1 7:53:27 2003	selected	Journal of Information Technology
dad@it.t...	193.92.233...	Tue Apr 1 9:12:12 2003	selected	Information Processing Letters
	155.207.12...	Tue Apr 1 10:03:45 2003		Data Mining and Knowledge Discovery
	155.207.12...	Tue Apr 1 10:09:19 2003		Searcher
	194.63.234.6	Tue Apr 1 10:18:46 2003		Teacher Education and Special Education
	155.207.11...	Tue Apr 1 3:32:02 2003		Army Lawyer, The

HEALlink log data analysis: association rules



Rule	Support	Confidence	▼ Lift	Absolute Support	Subtractive Lift	Items in Rule Body	Items in Rule Head	Group
[Mathematics]+[Science (General)]+[Technology (General)]+[History of Asia] ==> [History of scholarship and l...	7,3864%	100,0000%	13,54	26	0,93	4	1	1
[Mathematics]+[Science (General)]+[Technology (General)]+[History of Asia] ==> [Speculative philosophy]	6,2500%	84,6154%	10,64	22	0,77	4	1	1
[Mathematics]+[Science (General)]+[Technology (General)]+[History of Asia] ==> [History of education]	5,9659%	80,7692%	9,48	21	0,72	4	1	1
[Mathematics]+[Science (General)]+[Technology (General)]+[History of Asia] ==> [Physics]	6,8182%	92,3077%	3,78	24	0,68	4	1	1

HEALlink log data analysis: IBM IM Visualizer



Classification (Decision Tree) with WEKA

- Titanic dataset: the values of four categorical attributes (class, age, gender, survived) for each of the 2201 people on board the Titanic when it struck an iceberg and sank
- retrieved from **http://www.cs.toronto.edu/~delve/data/titanic**
- learners are asked to:
  - calculate (manually, without using any DM software) the attribute value that, when considered all by itself, plays the most deterministic role in telling whether a passenger has survived the accident
  - manual calculations to be carried out in accordance with: (a) the information gain (entropy) maximization algorithm, and (b) the affinity analysis (apriori) algorithm
  - double-check the result obtained by utilizing the WEKA decision tree (J48) algorithm
  - transform the titanic data input so that they can be fed into the IBM DB2 association rules mining algorithm and interpret/comment on the new result output

Technical Issues

- WEKA: open source, runs 'everywhere' (MS Windows, Linux, Mac OSX)
- IBM DB2 DWE 9.5 server on IBM System x hardware (virtual 64-bit MS-Windows 2003 server).
  - IBM software is available for educational use only, from the IBM Academic Initiative program
  - MS software is available for educational use only, from the Microsoft MSDN Academic Alliance program
- IBM DB2 DWE 9.5 client (MS Windows)
  - many students have reported problems when they tried to install in on MS-Windows Vista; need to thoroughly document the problems encountered and recommend corrective actions (possibly: in collaboration with Microsoft and/or IBM) in a FAQ section of the the DBTech EXT web portal

References

Dunham M.H., *Data Mining: Introductory and Advanced Topics*, Prentice Hall; 1st edition (2002)

Witten I.H., Frank E., *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufman; 2nd edition (2005)

Roiger R., Geatz M., Data Mining: A Tutorial Based Primer, Addison Wesley; 1st edition (2002)