KD Ubiq Summer School 2008
Behavioural Modelling of a Grid System

Michele Sebag
CNRS − INRIA − Université Paris-Sud
http://tao.lri.fr

March 8th, 2008

Overview

Autonomic Computing

- A booming field of applications
- Machine Learning and Data Mining for Systems

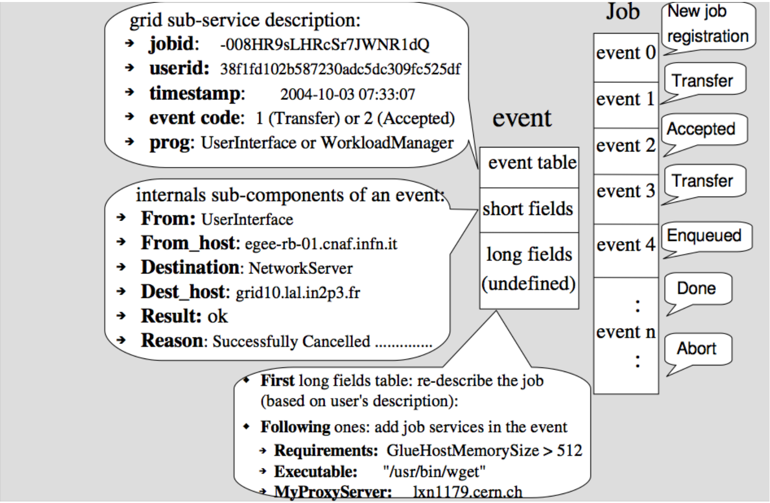
Autonomic Grid

- EGEE: Enabling Grids for e-Science in Europe
- Data acquisition, Logging and Bookkeeping files
- (change of) Representation, Dimensionality reduction

Modelling Jobs

- Exploratory Analysis and Clustering
- Clustering the jobs

Job representation



Xiangliang Zhang et al., ICDM wshop on Data streams, 2007

Job representation

Challenges

- Sparse representation, e.g. "user id"
- No natural distance

Prior knowledge

- Coarse job classification: succeeds (SUC) or fails (FAIL)
- Many failure types: Not Available Resources (NAR); User Aborted (ABU); Generic and non-Generic Error (GNG).
- Jobs are heterogeneous
  - Due to users (advanced or naive)
  - Due to virtual organizations (jobs in physics $\neq$ jobs in biology)
  - Due to time: grid load depends on the community activity

Feature extraction

Slicing data

to get rid of heterogeneity

- Split jobs per user: $U_i = \{$ jobs of i-th user $\}$
- Split jobs per week: $W_j = \{$ jobs launched in j-th week $\}$

Building features

- Each data slice: a supervised learning problem (discriminating *SUCC* from *FAIL*)

$$h : \mathcal{X} \mapsto \mathbb{R}$$

- Supervised Learning Algorithms:
  - Support Vector Machine — SVMLight
  - Optimization of AUC — ROGER

Feature Extraction, 2

New features

Define

$h_{u,i}$ hypothesis learned from data slice U_i

$U : \mathcal{X} \mapsto \mathbb{R}^{\#u}$

$$U(\mathbf{x}) = (h_{u,1}(\mathbf{x}), \ldots h_{u,\#u}(\mathbf{x}))$$

Symmetrically $h_{w,i}$ hypothesis learned from data slice W_i

$W : \mathcal{X} \mapsto \mathbb{R}^{\#w}$

$$W(\mathbf{x}) = (h_{w,1}(\mathbf{x}), \ldots h_{w,\#w}(\mathbf{x}))$$

Change of representation

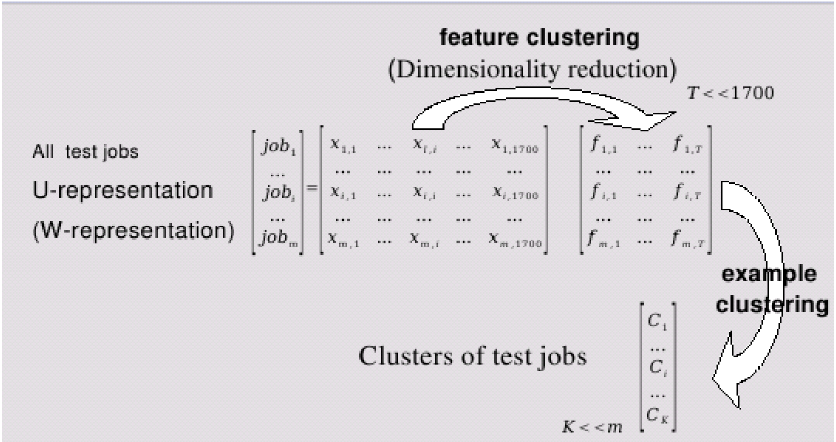
$$\begin{array}{lll} \mathcal{E} & \rightarrow & \mathcal{E}_U = \{(U(\mathbf{x}_i), y_i), i = 1 \ldots N\} \\ & \rightarrow & \mathcal{E}_W = \{(W(\mathbf{x}_i), y_i), i = 1 \ldots N\} \end{array}$$

Discussion

- Natural distance on $\mathbb{R}^d$
- But new attributes $h_{u,i}$ likely to be redundant

Feature Extraction: Double clustering

Slonim & Tishby, 2000



Experimental setting

The datasets

- Training set $\mathcal{E}$: 222,500 jobs 36% SUCC, 74% FAIL
- Test set $\mathcal{T}$: 21,512 jobs

Hypothesis construction

- SVM: one hypothesis per slice: $U : \mathcal{X} \mapsto \mathbb{R}^{34}$, $W : \mathcal{X} \mapsto \mathbb{R}^{45}$
- ROGER: 50 hypotheses per slice $U : \mathcal{X} \mapsto \mathbb{R}^{1700}$, $W : \mathcal{X} \mapsto \mathbb{R}^{2250}$

Clustering

Foreach $K = 5 \ldots 30$, Apply K-means to $\mathcal{T}$

- Considering new representations U and W
- Learned after SVM and Roger.

Goal of Experiments

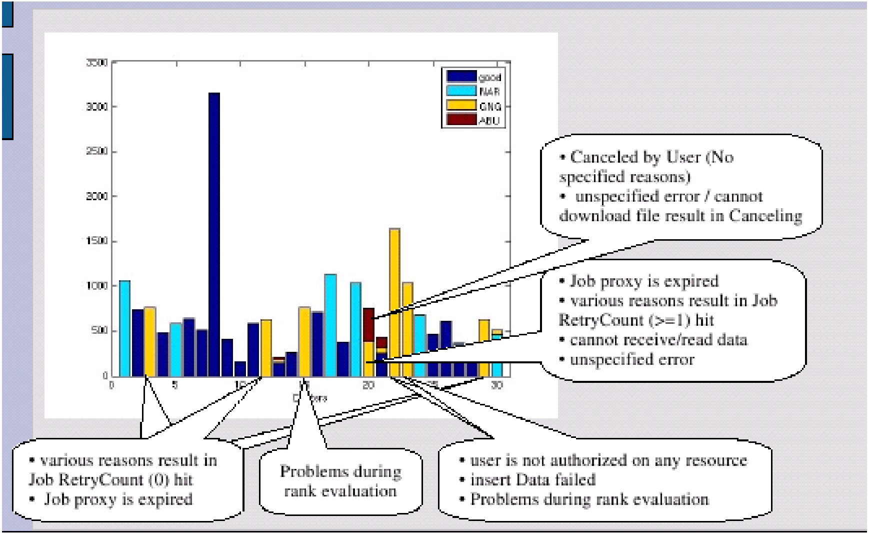
Interpretation

Examine the clusters

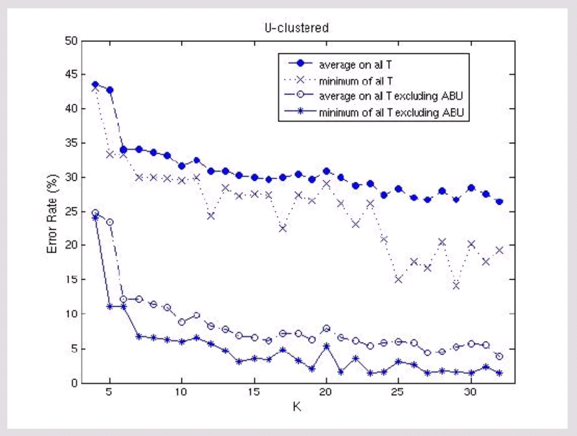
Stability

- Compare Δ_K and $\Delta_{K'}$
- Compare $\Delta_{K,U}$ and $\Delta_{K,W}$

Interpretation



Interpretation, 2



Interpretation, 3

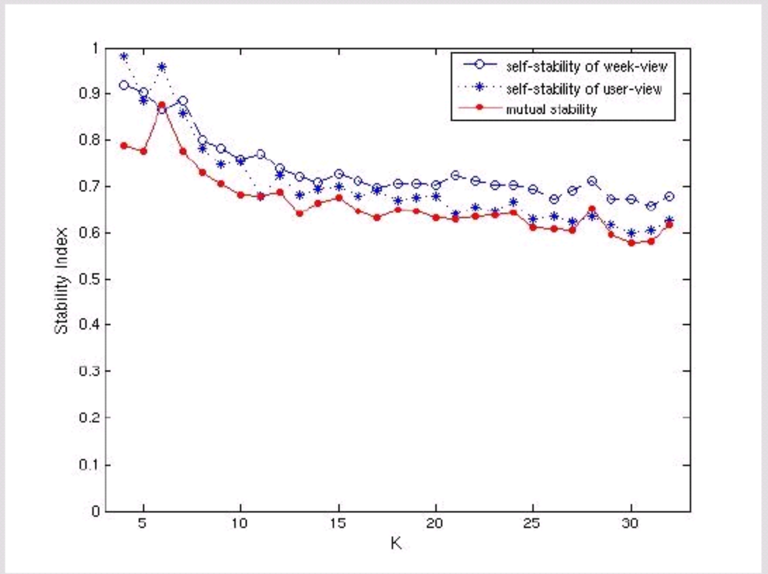
Pure clusters

- Most clusters are pure wrt sub-classes NAR, GNG which were unknown from the algorithm
- Finer-grained classes are discovered: Problem during rank evaluation; job proxy expired; insert Data failed
- ABU class (1.2%) is not properly identified: many reasons why job might be *Aborted by User*

Usage

Use prediction for user-friendly service
Anticipate job failures

Stability



self-stability of week-view
self-stability of user-view
mutual stability
Stability Index
K
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
5
10
15
20
25
30

Stability, 2

- Stability wrt initialization, for both W and U representations
- Stability of clusters based on W and U-based representations
- Decreases gracefully with K
  (optimal value $= 1$)

Grid Modelling, wrap-up

Conclusion

- Importance of representation as usual
- Clustering: stable wrt K and representation change
  re-discovers types of failures
  discovers finer-grained failures

Future work

- Cluster users (= sets of jobs)
- Cluster weeks (= sets of jobs)
- Find scenarios
  naive users gaining expertise;
  grid load & temporal regularities
- Identify communities of users.
- Use scenarios to test/optimize grid services (e.g. scheduler)

Autonomic Computing, wrap-up

Huge needs

- Modelling systems
  Black box to calibrate, train, optimize services
- Understanding systems — Hints to repair, re-design systems

Dealing with Complex Systems

- Findings often challenge conventional wisdom
- Theoretical *vs* Empirical models
- Complex systems are counter-intuitive — sometimes

Autonomic Computing, wrap-up, 2

Good practice

- No Magic !
  I don't see anything, I'll use ML or DM
- Use all of your prior knowledge
  If you can measure/model it, don't guess it!
- Have conjectures
- Test them!

Beware: False Discovery Rate

Thanks to

- Cécile Germain-Renaud
- Xiangliang Zhang
- Cal Loomis
- Nicolas Baskiotis
- Moises Goldszmidt
- The PASCAL Network of Excellence

  http://www.pascal-network.org

Borrowed slides

Pascal http://videolectures.net/

Clustering

- M. Meila. The uniqueness of a good optimum for K-means. ICML. 625-632,2006
- U. von Luxburg et al., Theoretical Foundations of Clustering, NIPS 2005

Classification

- SVMLight. T. Joachims. Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning. B. Scholkopf and C. Burges and A. Smola (ed.), MIT-Press, 41-56, 1999.
- ROGER. M. Sebag, N. Lucas, J. Azé. Impact studies and sensitivity analysis in medical data mining with ROC-based genetic learning. IEEE Int. Conf. on Data Mining, 637-640, 2003.