

Creating Knowledge from IT Events

Ira Cohen

HP-Labs

With: Michal Aharon, Gilad Barash,
Eli Mordechai, Arik Itskovic, Rafael
Dahan



© 2006 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice

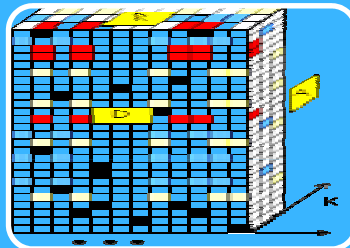


Technology Roadmap



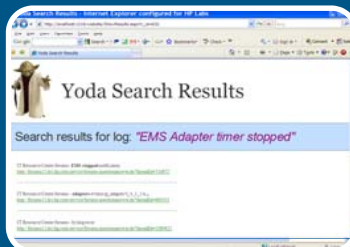
Create Dictionary of Event Types

- Transforms raw text logs to machine readable form
- Novel text clustering algorithm



Create Dictionary of Processes

- Group event types that characterize system behavior
- PARIS Algorithm: Principal Atom Recognition In Sets



Create Knowledge from Documents

- Search, rank and summarize external and internal sources
- Information association and quality

Event logs in raw form

12/1/2008 12:34:03 failed to retrieve the meta data of project 'null0' the session auth has failed.

12/1/2008 12:35:03 failed to get licenses for project session the session auth has failed.

12/1/2008 12:40:31 error processing request from 192.111.22.33 data starts with 0 \00000023\0 conststr

12/1/2008 12:44:03 unexpected failure while trying to ping user session #44444 the session auth has failed

12/1/2008 12:50:03 failed to retrieve the meta data of project 'null1' the session authentication has failed.

12/1/2008 12:50:05 unexpected failure while trying to ping user session #33333 the session auth has failed

12/1/2008 12:50:23 failed to get licenses for project session the session auth has failed.

12/1/2008 12:55:09 failed to get licenses for project session the session auth has failed.

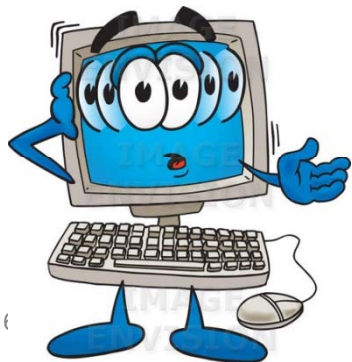
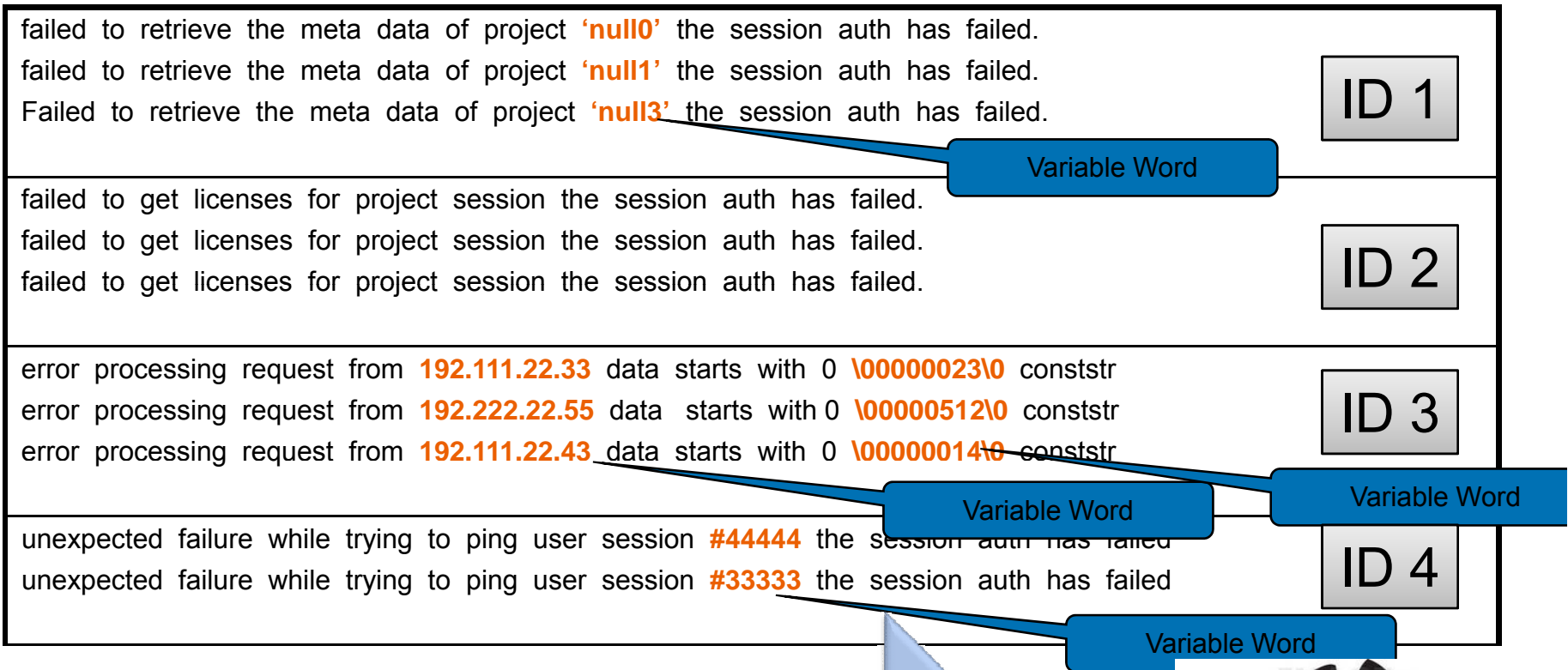
12/1/2008 12:56:22 error processing request from 192.222.22.55 data starts with 0 \00000014\0 conststr

12/1/2008 12:56:56 Failed to retrieve the meta data of project 'null3' the session auth has failed.

12/1/2008 12:57:03 error processing request from 193.111.26.33 data starts with 0 \00000512\0 conststr

12/1/2008 12:57:25 error processing request from 192.111.22.43 data starts with 0 \00000014\0 conststr

Let's Rearrange the Messages...



Template Discovery:
Assign IDs to Event Types
4 templates



Requirements for Template Discovery

1. Online

- Produce immediate value

2. Consistent

- Template assignment of a message should remain consistent over time

3. Efficient

- Keep up with incoming message rates

Template Discovery Algorithm:

Incremental Text Clustering

- Step 1: “Rough” clustering:
 - Creating/Assigning events to root clusters
- Step 2: Cluster refinement:
 - Splitting root clusters

Output: Forest of clusters

Template discovery algorithm

$$D_{\cos}(A, B) = \frac{\text{Match}(a_i, b_j)}{\sqrt{|A| \cdot |B|}}$$

Min Similarity
Threshold: 0.6

=0.83

Clustering example:

m1: B C D F A B

m2: B C D F A B J

m3: A C D F E K

m4: B C D F **E** B

1000 appearances of m4

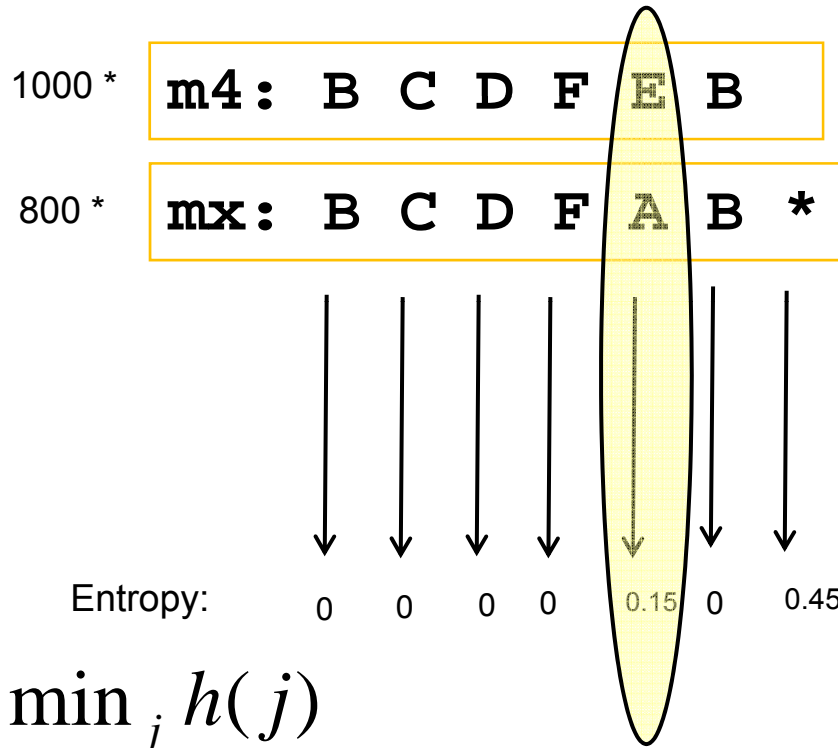
800 appearances of BCDFAB_



m3



Entropy Calculation for Split



$$h(j) = -\sum_{k=1}^n P_{kj} \bullet \log(P_{kj})$$

$$\arg \min_j h(j)$$

where $\varepsilon < h(j) < threshold$

Template discovery algorithm

$$D_{\cos}(A, B) = \frac{\text{Match}(a_i, b_j)}{\sqrt{|A| \cdot |B|}}$$

=0.66

Clustering example:

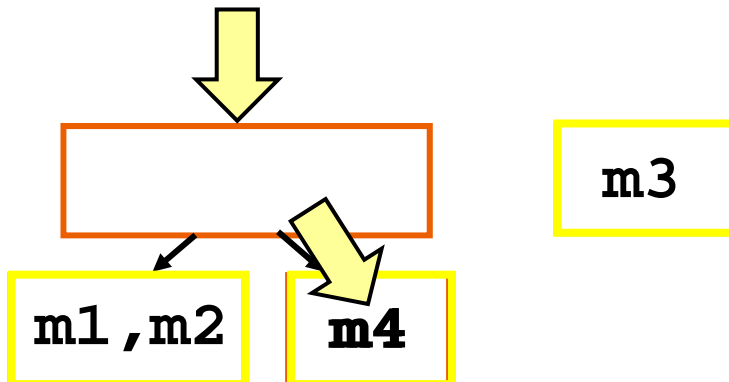
m1: B C D F A B

m2: B C D F A B J

m3: A C D F E K

m4: B C D F **E** B

mxx: B C D F E D



Results

- Datasets

Source	Number of events	Number of distinct events
Business App 1	4,210,513	153,619
Printer Press	11,204	5,631
Windows Events	66,102	25,340
Business App 2	483,768	70,102

Results

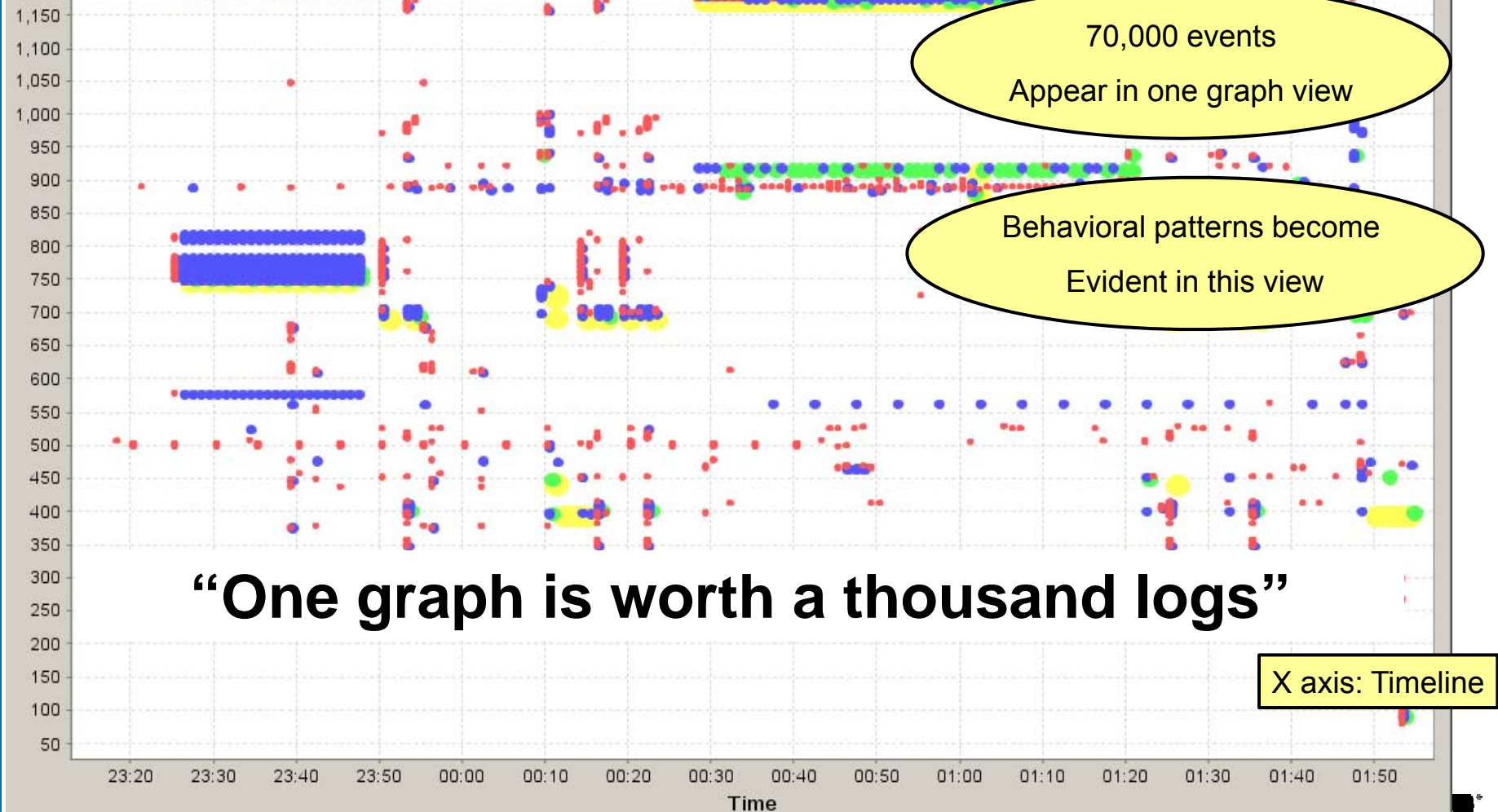
- Template identification

Source	Number of events	Number of distinct events	Number of clusters (templates)
Business App 1	4,210,513	153,619	4,193
Printer Press	11,204	5,631	204
Windows Events	66,102	25,340	476
Business App 2	483,768	70,102	1,115

Representation Accuracy: 95%

Visualizing the logs: Business App 2 Event Timeline

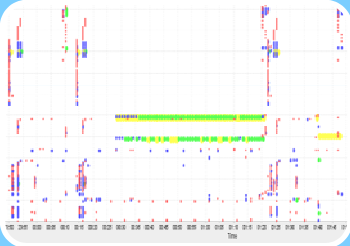
Y axis: msg ID



“One graph is worth a thousand logs”

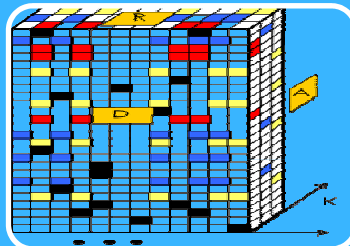
X axis: Timeline

Technology & Innovation Roadmap



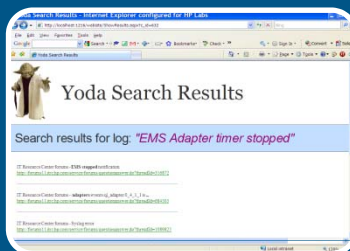
Create Dictionary of Event Types

- Transforms raw text logs to machine readable form
- Novel text clustering algorithm



Create Dictionary of Processes

- Group event types that characterize system behavior
- PARIS Algorithm: Principal Atom Recognition In Sets



Create Knowledge from Documents

- Search, rank and summarize external and internal sources
- Information association and quality

The Problem

Discovering Process Patterns

Database Connection Startup

- (1) JDBC3 getGeneratedKeys(): disabled
- (3) Connection release mode: auto
- (5) getConnectionURLs=tcp://websiteURL:2507
- (6) Query translator: hql.ast.ASTQueryTranslatorFactory
- (9) create connection. connectId
- (10) mercury_db_loader_DB_Loader user=;pwd=;

Service Manager startup

- (2) SH remote was null. Exported object monitor.
- (4) Add task Main Flow
- (7) Register provider class dataentry.loader.LoaderMain
- (8) Service manager started
- (3) Connection release mode: auto
- (11) mercury_db_loader is up and running

Optional messages

Same message
In different process

- (1) JDBC3 getGeneratedKeys(): disabled
- (2) SH remote was null. Exported object monitor.
- (3) Connection release mode: auto
- (4) Add task Main Flow

Challenges:

1. Pattern Interleaving

2. Noisy events

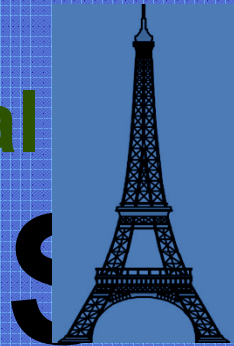
3. Non-determinism

4. 1-to-many mapping between event and process.

- (5) getConnectionURLs=tcp://websiteURL:2507
- (6) Query translator: hql.ast.ASTQueryTranslatorFactory
- (7) Register provider class dataentry.loader.LoaderMain
- (8) Service manager started
- (9) create connection. connectId
- (10) mercury_db_loader_DB_Loader user=;pwd=;
- (3) Connection release mode: auto
- (11) mercury_db_loader is up and running

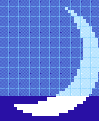
P rincipal

I n



tom

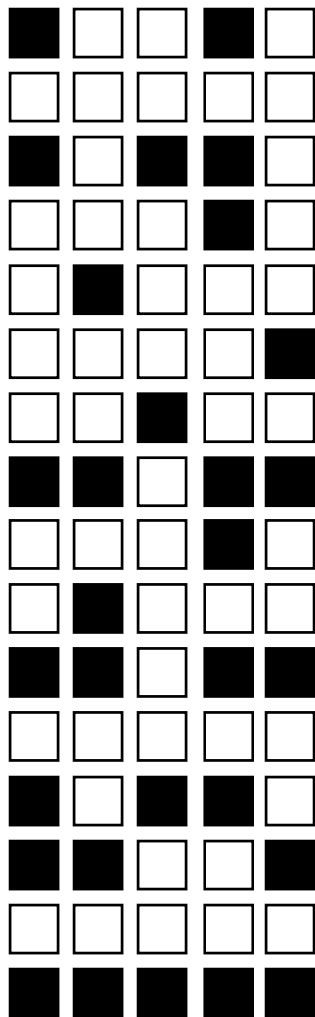
R ecognition



PARIS

Data:

$D_1 D_2 D_3 D_4 D_5$



D_N



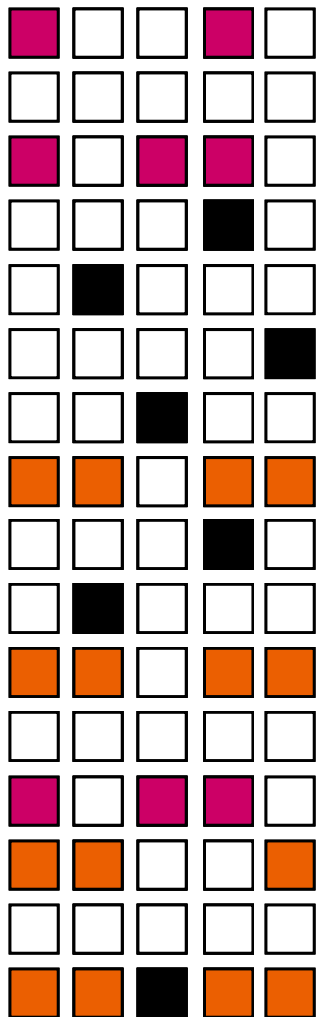
...

- Gets as input a large number of sets, that are assumed to have some mutual characterization.

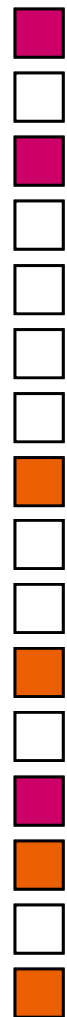
PARIS

Data:

$D_1 D_2 D_3 D_4 D_5$

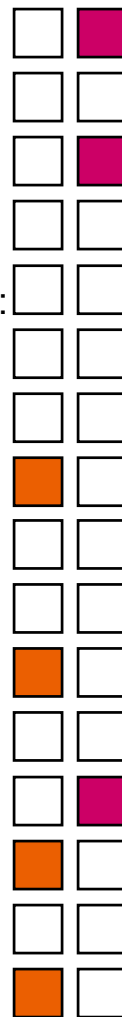


D_N



...

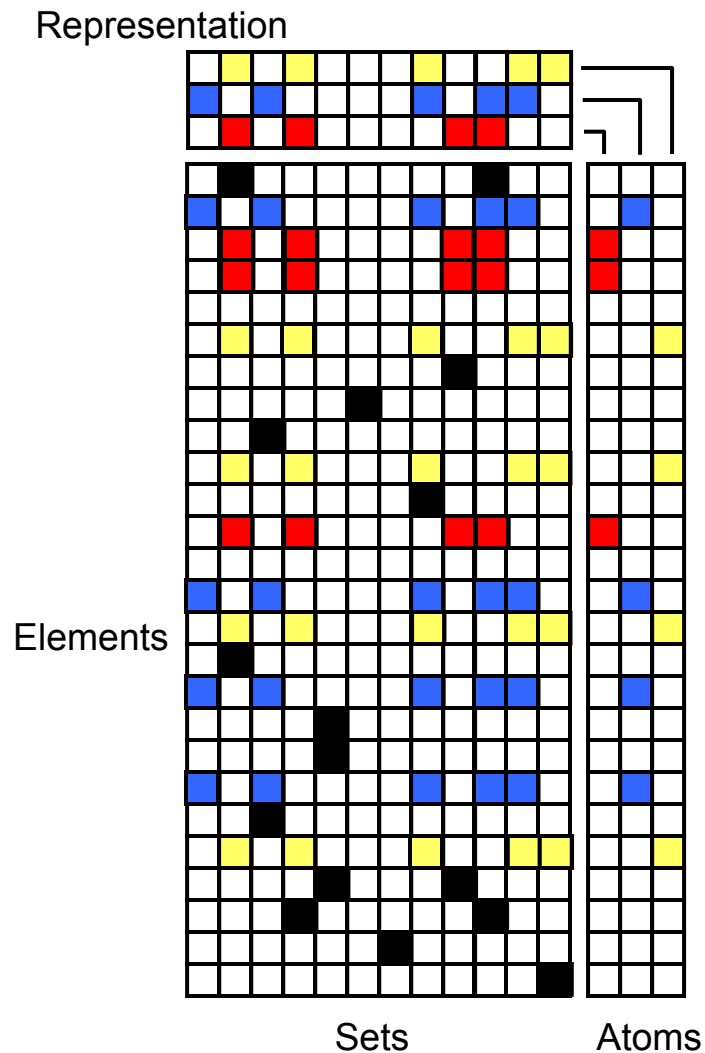
$A_1 A_2$



Atoms:

- Gets as input a large number of sets, that are assumed to have some mutual characterization.
- Detect principal sets of elements that tend to appear together in the data.
- Overcome non-exact repetitions
- Ignore additional noise

PARIS: Requirements



- Representation error must be small, but not necessarily zero.
- Representation should serve some sense of compression of the data (sparsity).
- Minimal number of atoms (K).

PARIS Cost Function

Minimize the representation error of the data.

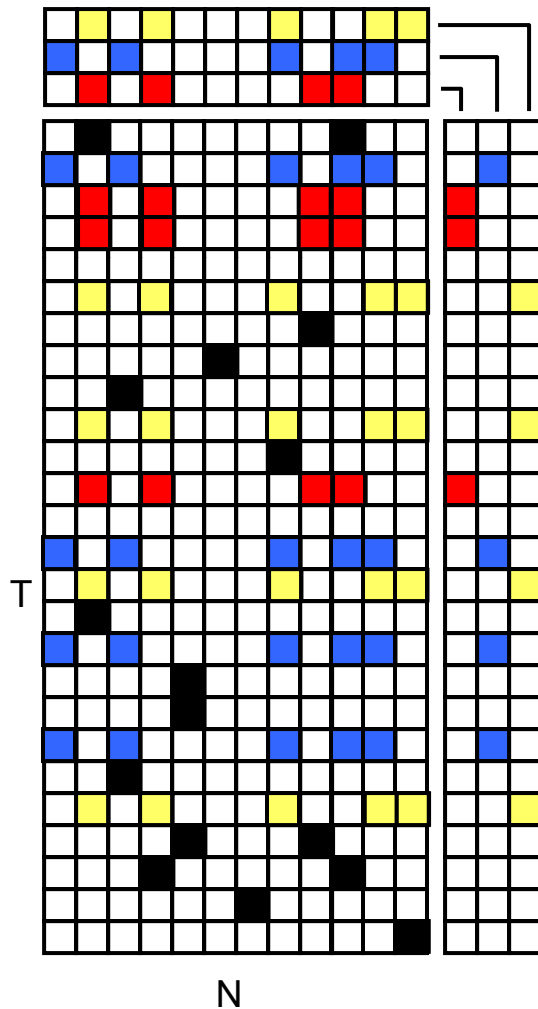
$$PCF = \arg \min_{A,R} \left(\sum_{i=1}^N d_r(D_i, R(A, R_i)) \right) + \left(\sum_{i=1}^N \mu_i |R_i| \right) + (\tau |A|)$$

Minimize the size of the representation (compression).

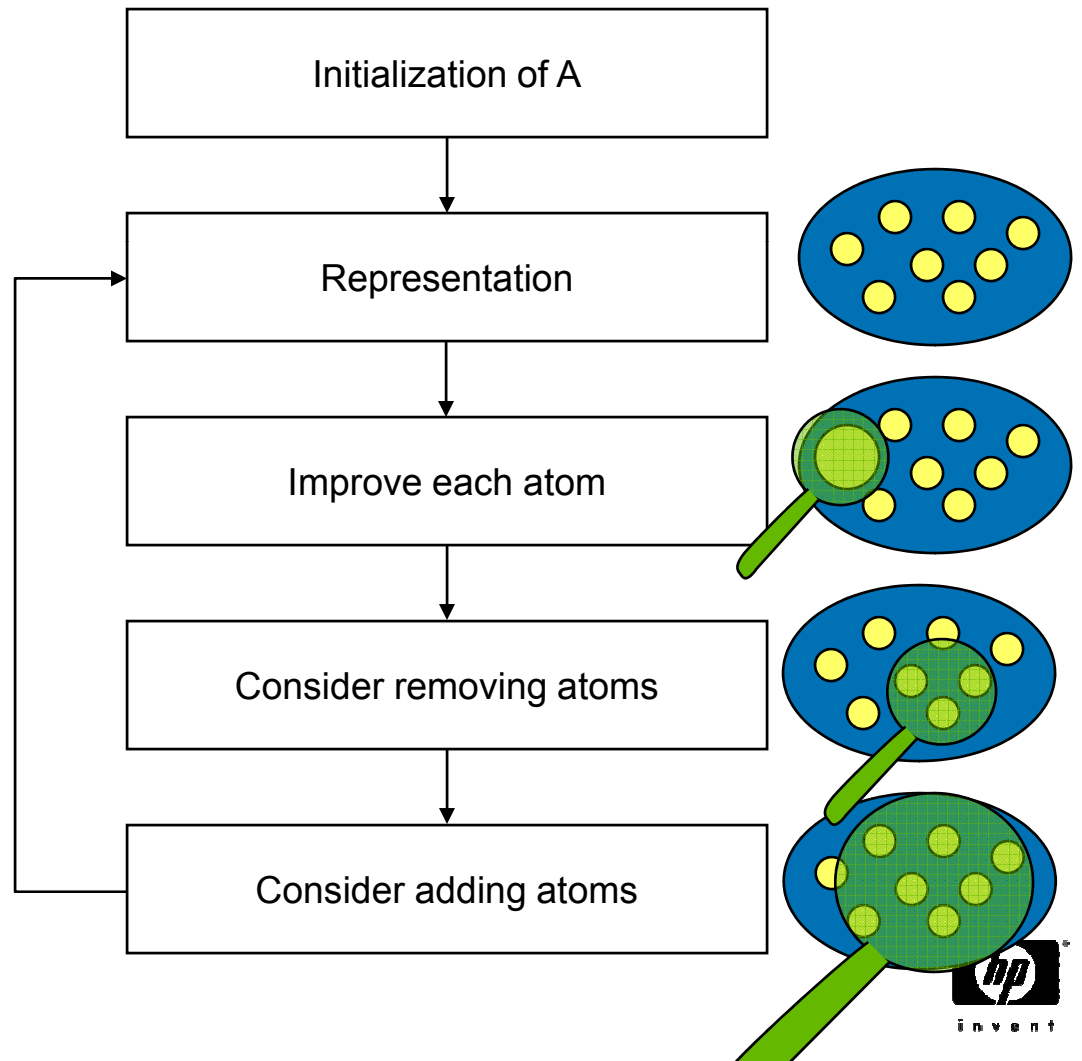
Minimize the number of principal atoms.

- Representation error must be small, but not necessarily zero.
- Representation should serve some sense of compression of the data (sparsity).
- Minimal number of atoms (K).

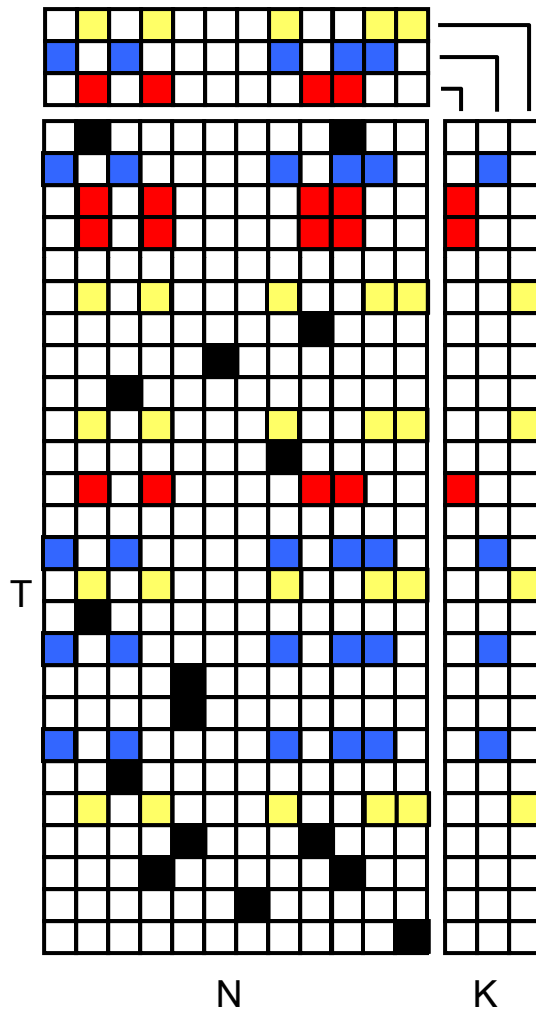
PARIS algorithm



K

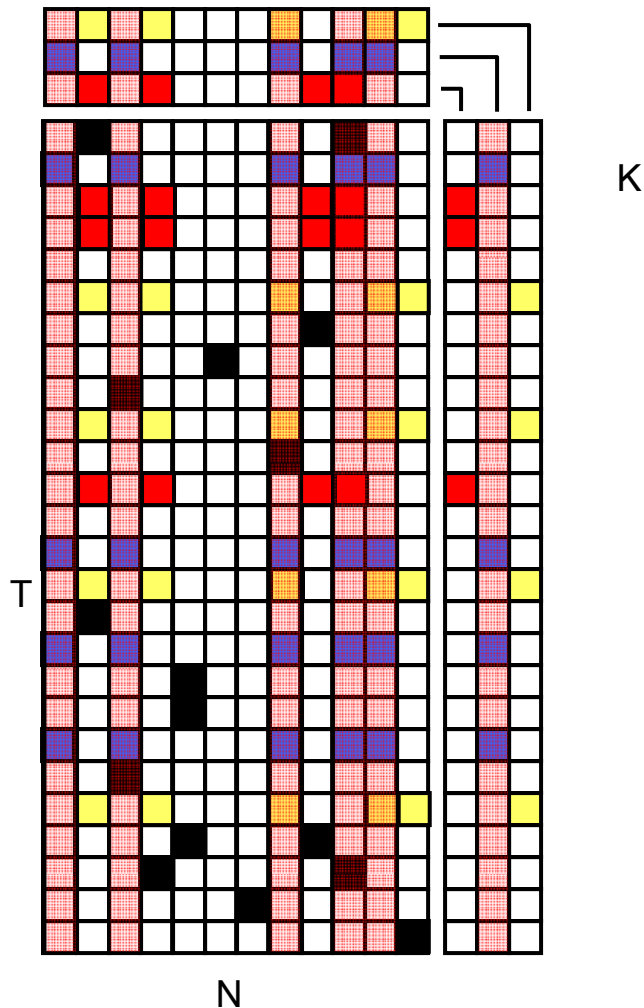


Representation



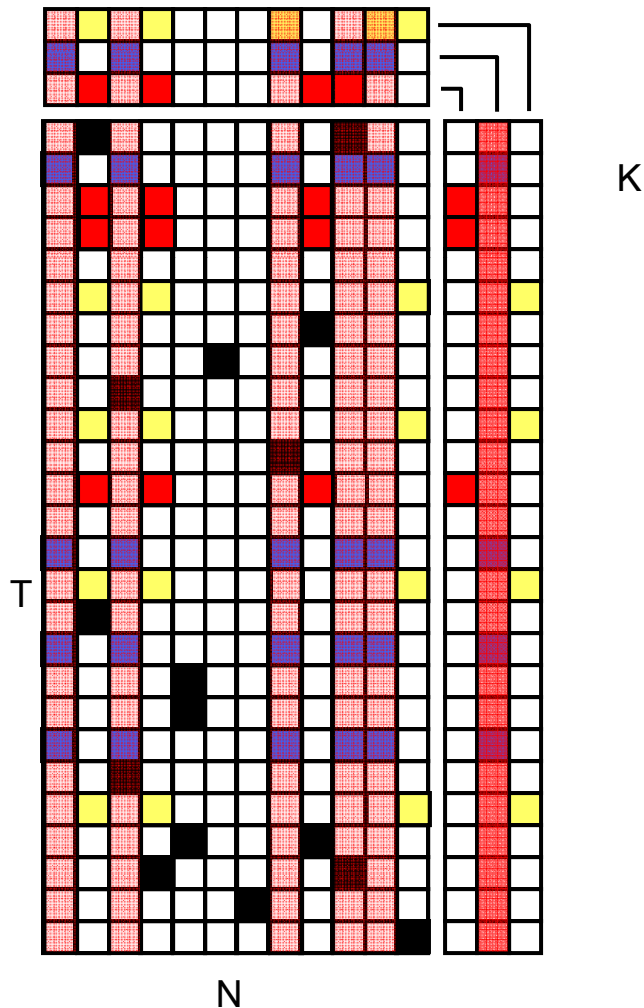
- Fixing A , we find a representation for each data set D_i .
- Problem provably NP-Hard
- Solution: Greedy algorithm

Single Atom Improvement



- For each $1 \leq j \leq K$, we fix all other atoms and update \mathbf{A}_j and its relevant part in the representation.

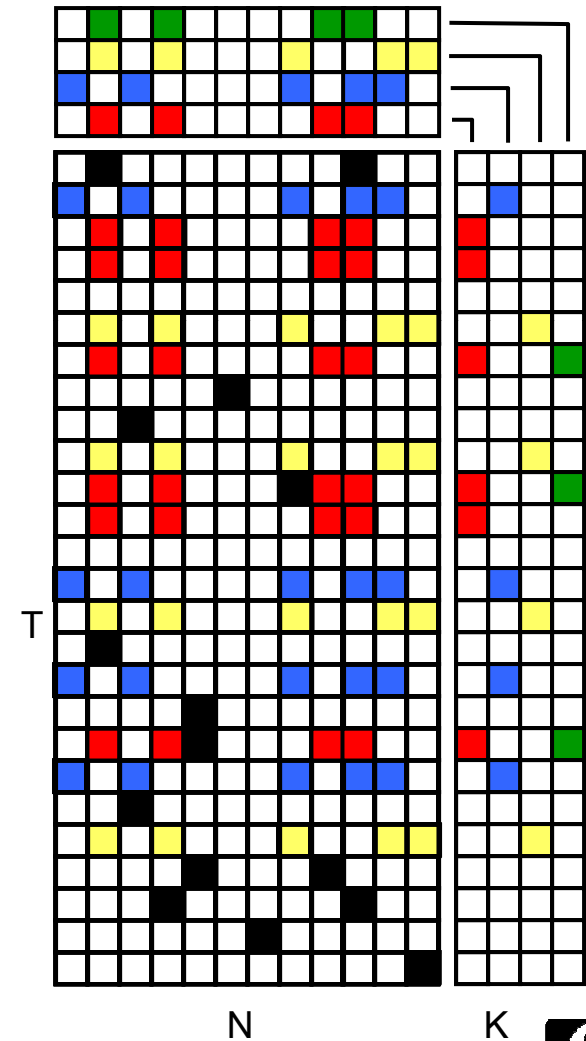
Single Atom Improvement



- For each $1 \leq j \leq K$, fix all other atoms and update \mathbf{A}_j and its relevant part in the representation.

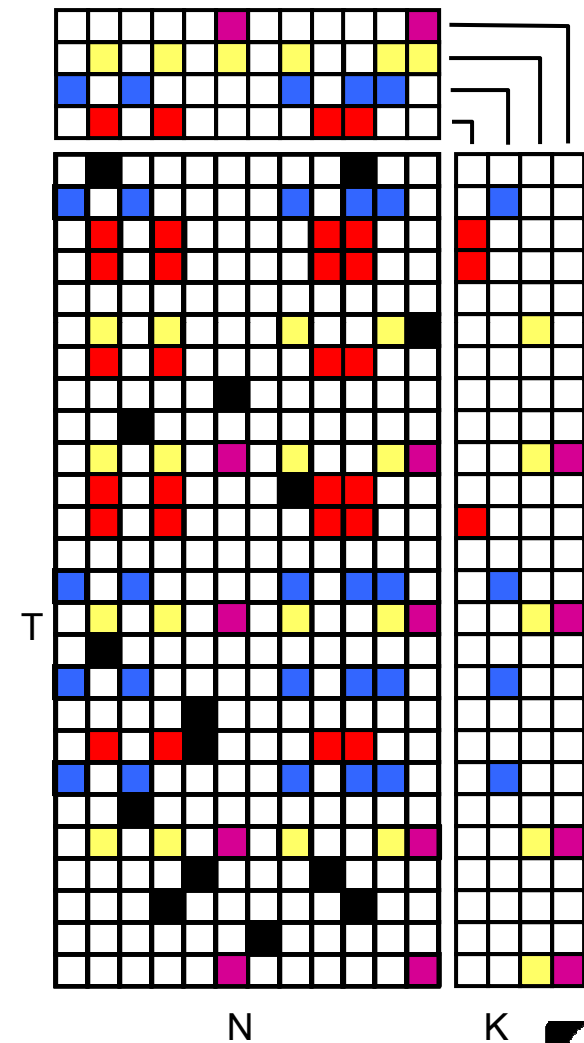
Removing Atoms

- We consider small sets of atoms and consider union of atoms or removing atoms in the following cases:
 - A set of atoms that tend to appear together in most representations can be united .
 -



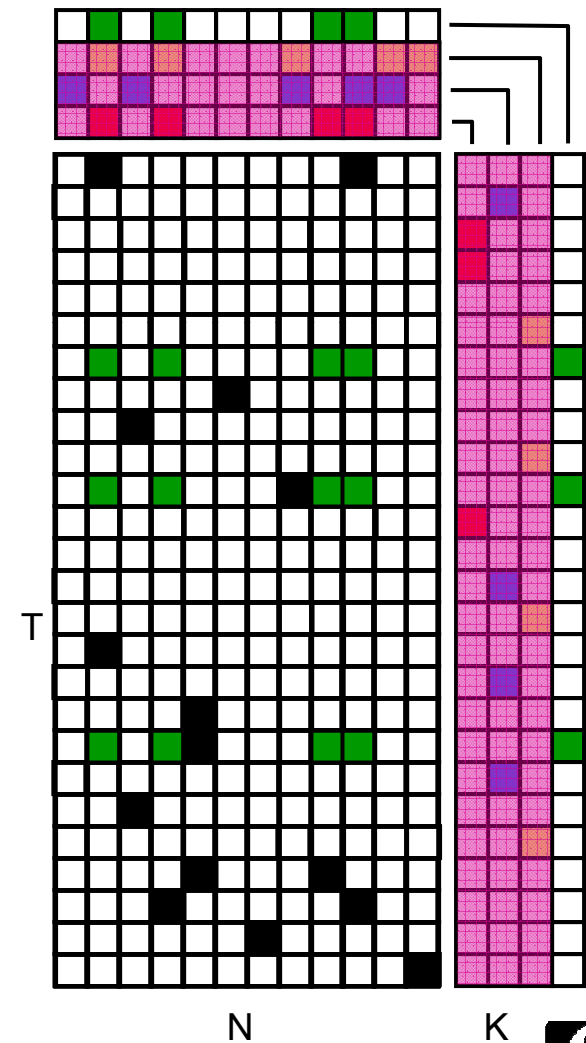
Omitting Atoms

- We consider small sets of atoms and consider union of atoms or dismiss atoms in the following cases:
 - A set of atoms that tend to appear together in most representations can be united .
 - A set of atoms that share many common elements and represent distinct sets of data sets can be united.



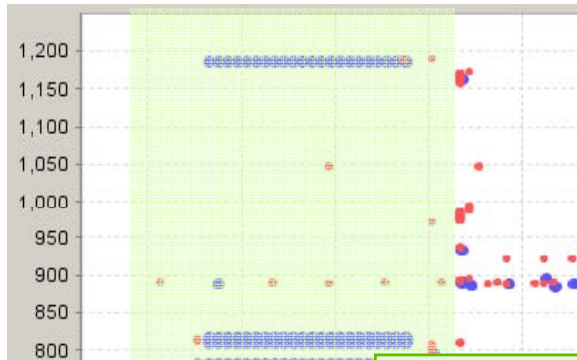
Adding atoms

- We detect regularities in the overall representation error.
- New atoms are designed to represent these regularities.



PARIS Result: Correct Process Identification

Buss App 2 Logs



Atom ID: 27

- 734 User operation - stop nanny
- 748 message_broker STOPPED
- 753 Input main(String[] args):
- 754 Going to call WrapperManager.start(new Main(), args)
- 755 Initializing Spring files
- 757 Path for spring files is E:\HPBAC\conf\supervisor\spring
- 759 Loading spring file
- 764 NannyConfig

Application Failure State

Atom ID: 12

- 890 Failed creating SiS sample
- 924 Failed processing http request report_ss_samples, from remote
Failed to acquire lock for publ
- 1183 Failed processing http request report_transaction, from remote
Failed to acquire lock for publ

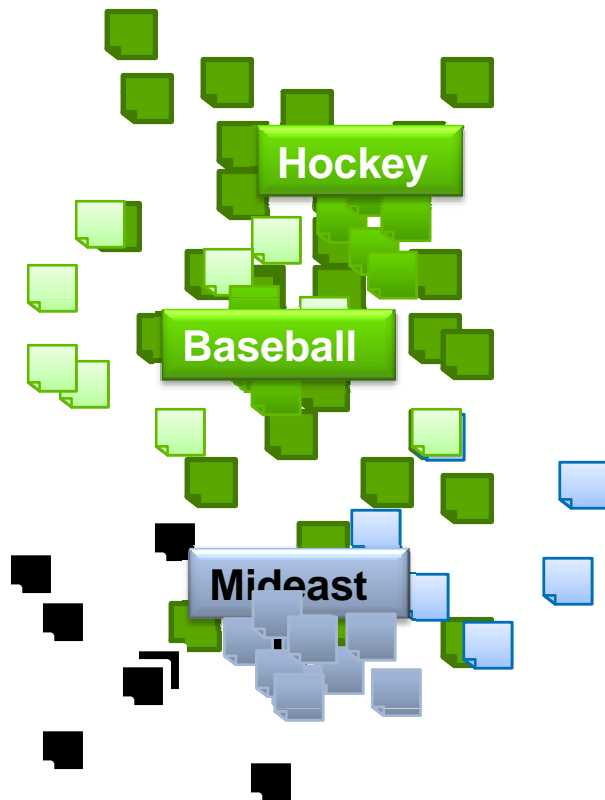
Service Restart

- 768 Autodetecting user-defined JMX MBeans
- 769 Bean with name 'nannyManager' has been autodetected for JMX exposure
- 770 HTTP adapter port is 11021
- 771 Succeeded adding html adapter
- 772 manager thread loop started.
- 773 Verifying time diff between cpp (local machine) and Java.
- 774 Log file of time diff is: E:\HPBAC\tools\TimeDiff\time_diff.log
- 776 Run java time diff
- 780 Trying to initialize Properties Manager
- 792 Config server check passed
- 793 Prerequisites have been met
- 794 start() Nanny Manager
- 795 Nanny Manager need to start all services?:true
- 796 Going to start all services.
- ...

PARIS Result: Document Representation

Newsgroup data set:

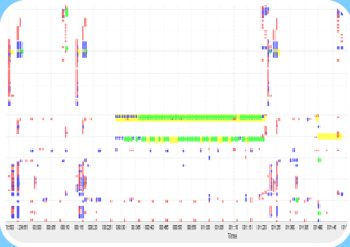
Document corpus labeled to 20 topics



Atoms

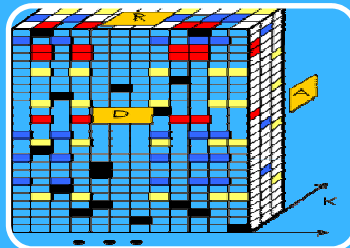


Technology & Innovation Roadmap



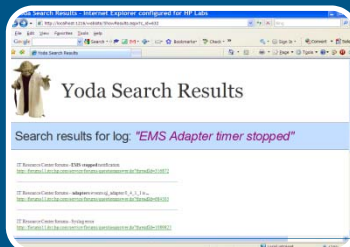
Create Dictionary of Event Types

- Transforms raw text logs to machine readable form
- Novel text clustering algorithm



Create Dictionary of Processes

- Group event types that characterize system behavior
- PARIS Algorithm: Principal Atom Recognition In Sets

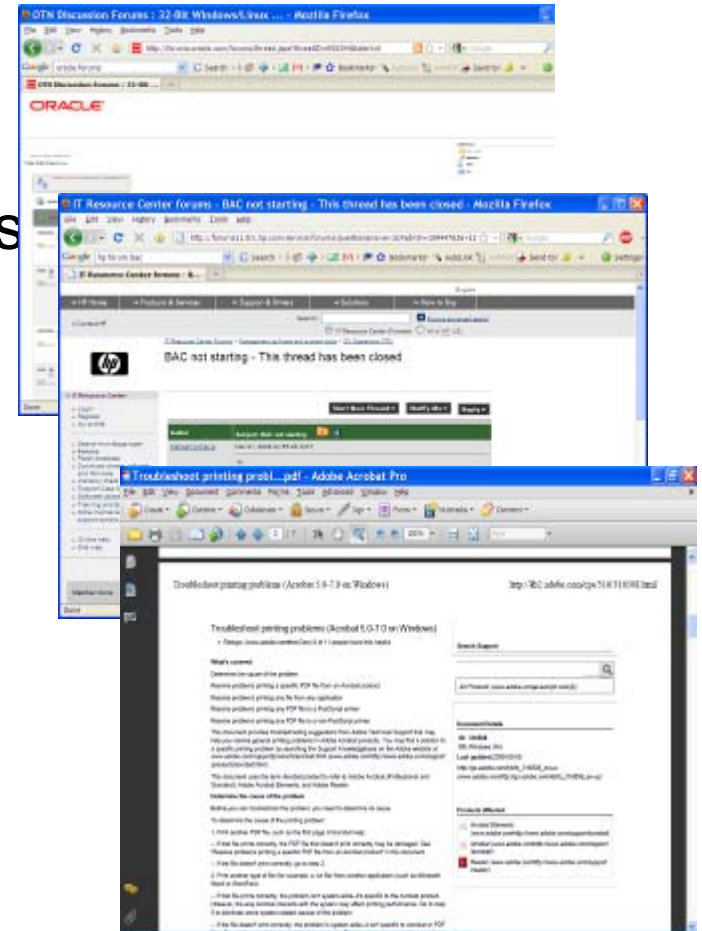


Create Knowledge from Documents

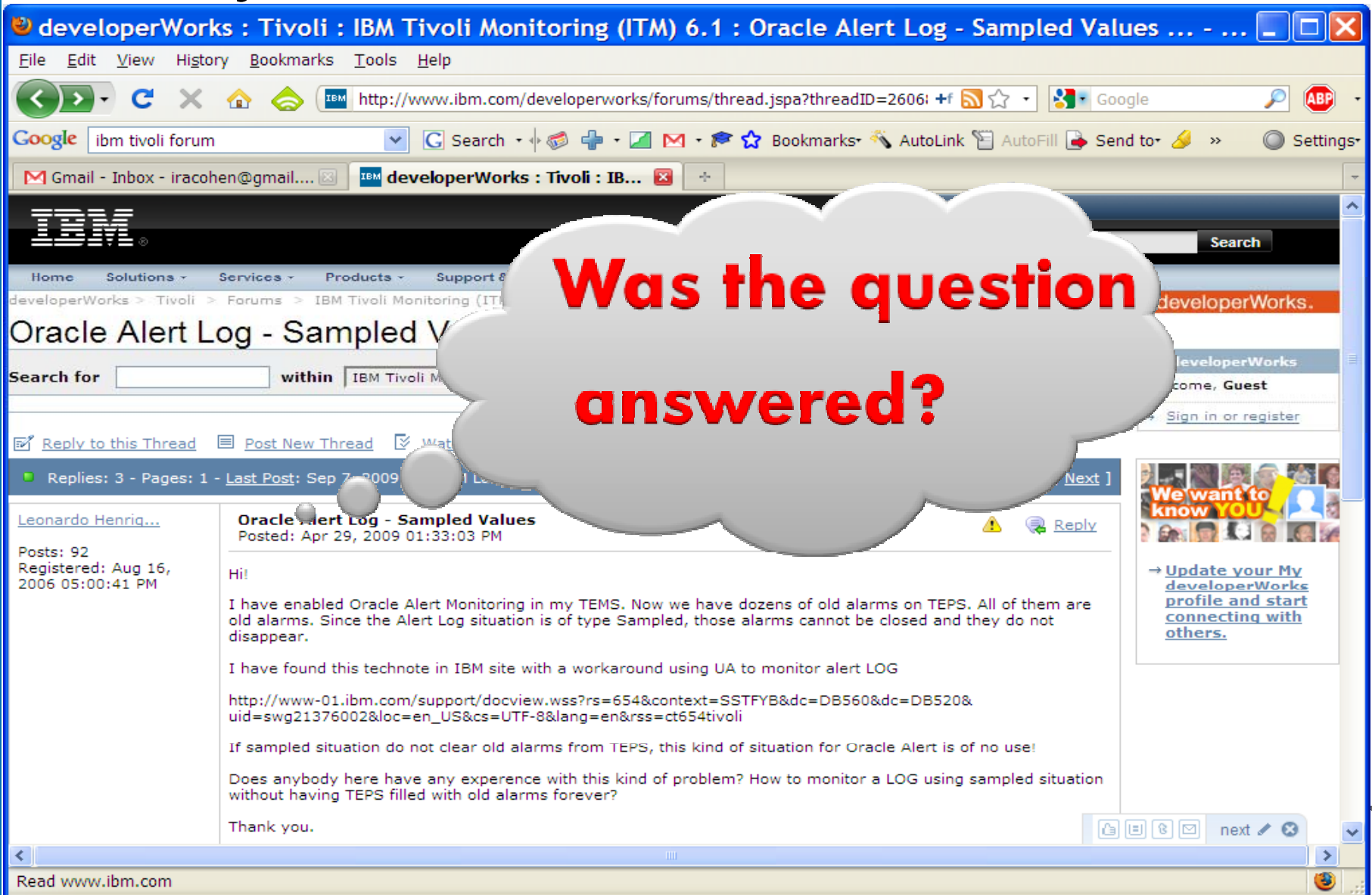
- Search, rank and summarize external and internal sources
- Information association and quality

Information Extraction from documents/wikis/forums

- Uses:
 - Link to events and problem periods
 - Create knowledge on problem types
 - Extract resolutions
- Innovation:
 - Relevancy of knowledge sources
 - Quality of information
 - Concept extraction
 - Document clustering



Quality of Information



The image shows a screenshot of a web browser displaying a forum thread on IBM developerWorks. The browser's address bar shows the URL: <http://www.ibm.com/developerworks/forums/thread.jspa?threadID=2606>. The forum page title is "Oracle Alert Log - Sampled Values". A large, white, cloud-shaped graphic with a red border is overlaid on the page, containing the text "Was the question answered?".

The forum post is by Leonardo Henriquez, posted on Apr 29, 2009, 01:33:03 PM. The post content is as follows:

Hi!

I have enabled Oracle Alert Monitoring in my TEMS. Now we have dozens of old alarms on TEPS. All of them are old alarms. Since the Alert Log situation is of type Sampled, those alarms cannot be closed and they do not disappear.

I have found this technote in IBM site with a workaround using UA to monitor alert LOG

http://www-01.ibm.com/support/docview.wss?rs=654&context=SSTFYB&dc=DB560&dc=DB520&uid=swg21376002&loc=en_US&cs=UTF-8&lang=en&rss=ct654tivoli

If sampled situation do not clear old alarms from TEPS, this kind of situation for Oracle Alert is of no use!

Does anybody here have any experience with this kind of problem? How to monitor a LOG using sampled situation without having TEPS filled with old alarms forever?

Thank you.

Was the question answered? : Results

Extract

- Collected 5500 Oracle forum threads, 1300 IBM forum threads
- Extracted 10 features



Train

- Training classifiers on threads from one domain, testing on the other

Classif y

Train/Test	Oracle	IBM
Oracle	90%	85%
IBM	79%	97%

Summary

- Presented system for creating knowledge from events
- Exploring uses in other domains, e.g., PARIS for collaborative filtering
- System currently being tested in various IT environments
- Publications available (ECML'09, HP-Labs Tech reports)

Thank you

Q&A

