Text Classification on a Grid Environment

Valeriana G. Roncero – valery@nacad.ufrj.br
Myrian C. A. Costa – myrian@nacad.ufrj.br
Nelson F. F. Ebecken – nelson@ntt.ufrj.br

COPPE/Federal University of Rio de Janeiro
Brasil

NACAD
High Performance Computing Center

ntt
Núcleo de Transferência de Tecnologia - UFRJ

VECPAR'10
COPPE
UFRJ

Agenda

- Objective
- Aîuri Portal
- Text Mining Proccess
- Semi-supervised Learning
- Naive Bayes Classification
- Expectation-Maximization Task
- Categorization Algorithm
- Grid Environment
- Testing
- Results
- Conclusions

COPPE UFRJ

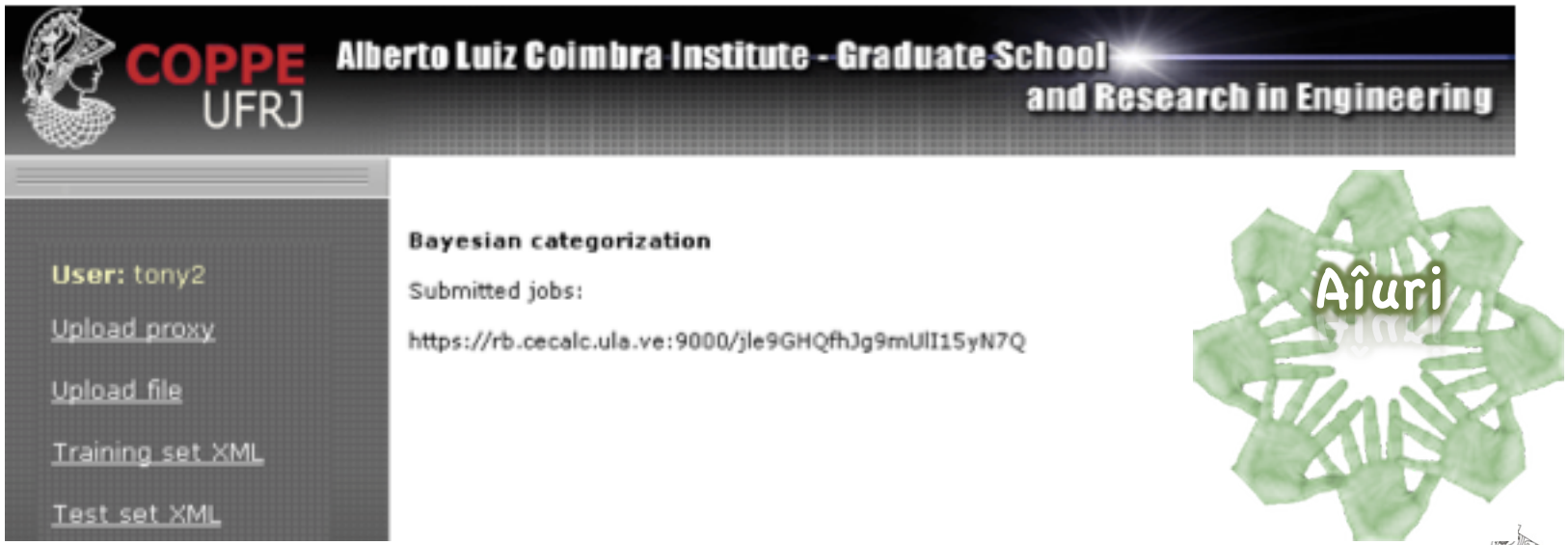
Objective

- Expand the functionality of **Aîuri Portal** with a **grid service** which implements a **semi-supervised learning** algorithm to categorize a collection of documents from labeled and unlabeled ones.

COPPE UFRJ

Aîuri Portal

– A high performance collaborative academic environment for education and research on computational intelligence, which permits the use of data/text/web mining tasks, across multiple grid environments and other execution.



Aîuri Portal

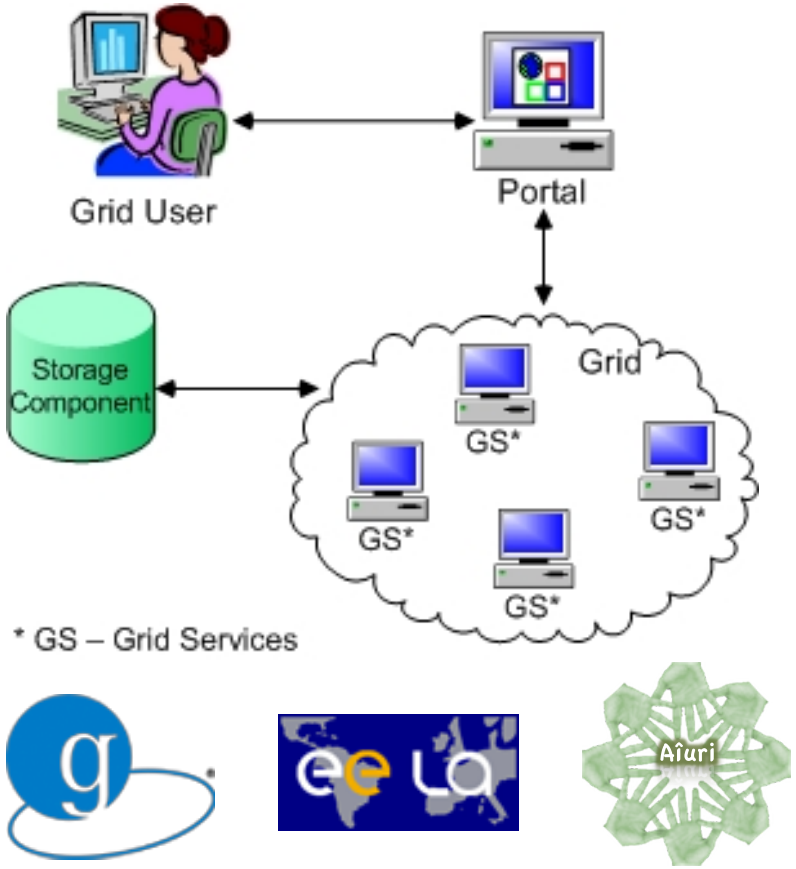
- Aîuri = mutual help

- Development of a framework which includes a user interface, data/text mining tasks, database access and a visualization tool;
- Utilization of open source components;
- Implementation and testing of new algorithms.

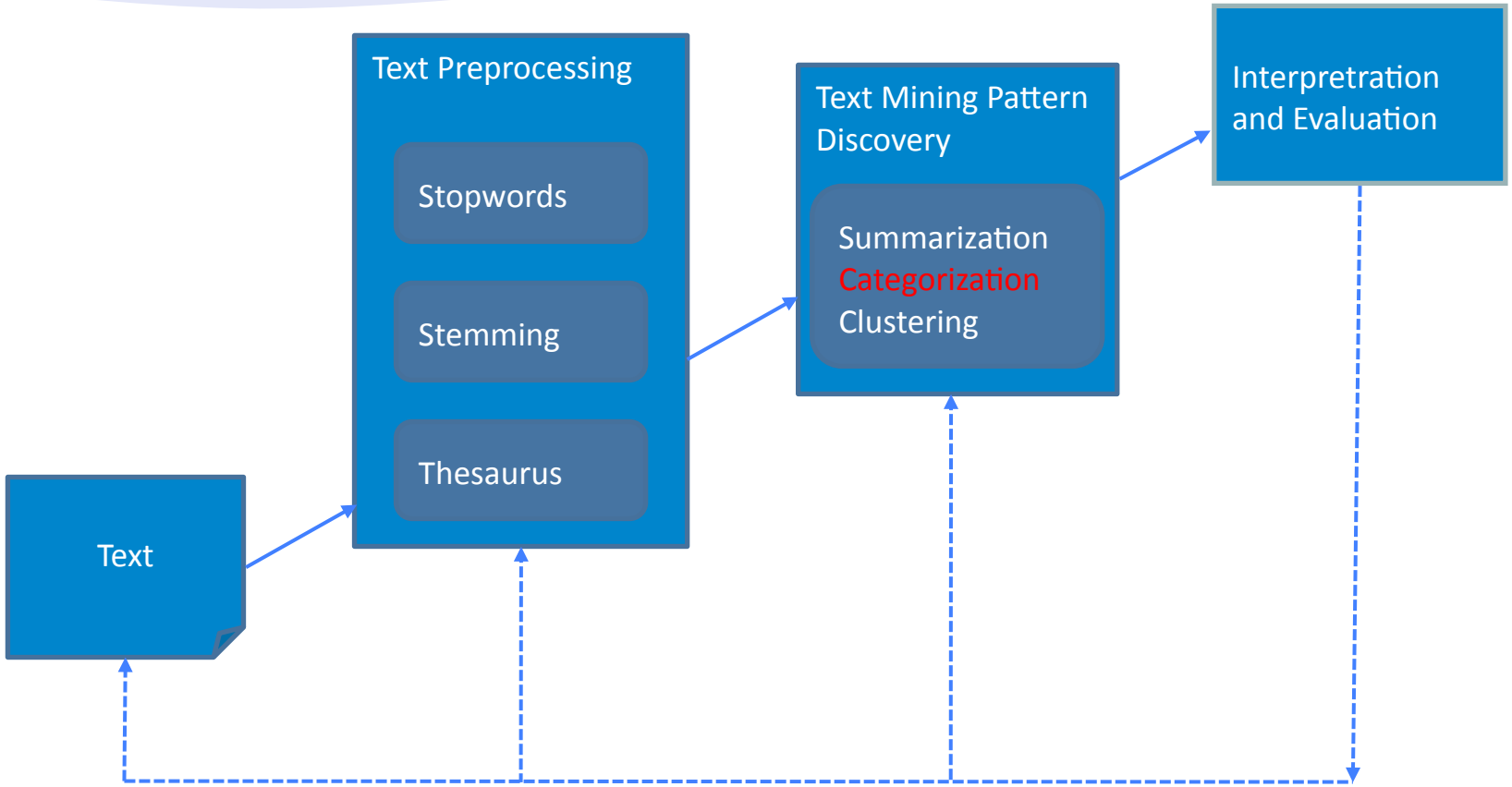
COPPE UFRJ

Aîuri Portal

- Aîuri Components:
  - **Portal**: interface between user and the services
  - **Grid services**: data and text mining tasks
  - Storage: maintain all the archives uploaded to the environment



Text Mining Process



Semi-supervised Learning

- Difficulty to obtain large sets of labeled documents to classify.
- Using the Expectation Maximization algorithm, the text classifier can learn using a small number of labeled documents with a large pool of unlabeled documents.
- Grid environments provide the tools to implement and execute distributed Text Mining tasks.

COPPE UFRJ

Semi-supervised Learning

- Supervised Learning
  - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations.
- Semi-supervised Learning
  - Use of both labeled and unlabeled data for training, typically a small amount of labeled data with a large amount of unlabeled data.
  - The Expectation Maximization algorithm with Naive Bayes for text documents classification is reported to provide improvements in classification (Nigam et. al., 2000).

COPPE UFRJ

Naive Bayes Classification

- A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with independent feature model.

$$P(c_k|D) = P(c_k) \times \frac{\prod_{j=1}^{m} P(d_j|c_k)}{P(D)}$$

where,

- D is a set of text documents: $D = \{d_1, d_2, \ldots, d_m\}$
- c_k is a possible class from a set of predefined classes: $C = \{c_1, c_2, \ldots, c_n\}$
- $P(c_k)$ is the class probability and can be estimated from training data

Naive Bayes Text Classification

- " v_{NB} is the classification that maximizes the probability of observing the words that were actually found in a document, subject to the usual naive Bayes independence assumption."

$$v_{NB} = \arg\max_{c_k \in C} P(c_k) \prod_i P(a_i | c_k)$$

  - a_i describes the attributes probabilities
  - $P(a_i=w_j|c_k)$ can be estimated based on the fraction of each class of a document

Expectation-Maximization Task

- Iterative procedure to estimate the maximum likelihood hypothesis values of parameters in models.
- Consists in two steps: E-step and M-step
- E-step : Expectation
  - Calculate the probable labels for the unlabeled documents using the observed text documents and current parameter estimated.
- M-step: Maximization
  - Compute the new maximum likelihood hypothesis values using the probable labels estimated on the Expectation step

COPPE UFRJ

Expectation-Maximization Task

- Given:
  - Labeled data $\mathcal{X} = \{x_1, \ldots, x_m\}$ and Unlabeled data $\mathcal{Z} = \{z_1, \ldots, z_m\}$
  - $\mathcal{Y}$ is the full data $\mathcal{Y} = \mathcal{X} \cup \mathcal{Z}$
  - h are the current hipothesized values
- Determine:
  - h' that (locally) maximizes $E[\ln P(\mathcal{Y}|h')]$
- Define:
  - likelihood function $Q(h', h)$ that gives $E[\ln P(\mathcal{Y}|h')]$ as a function of h'

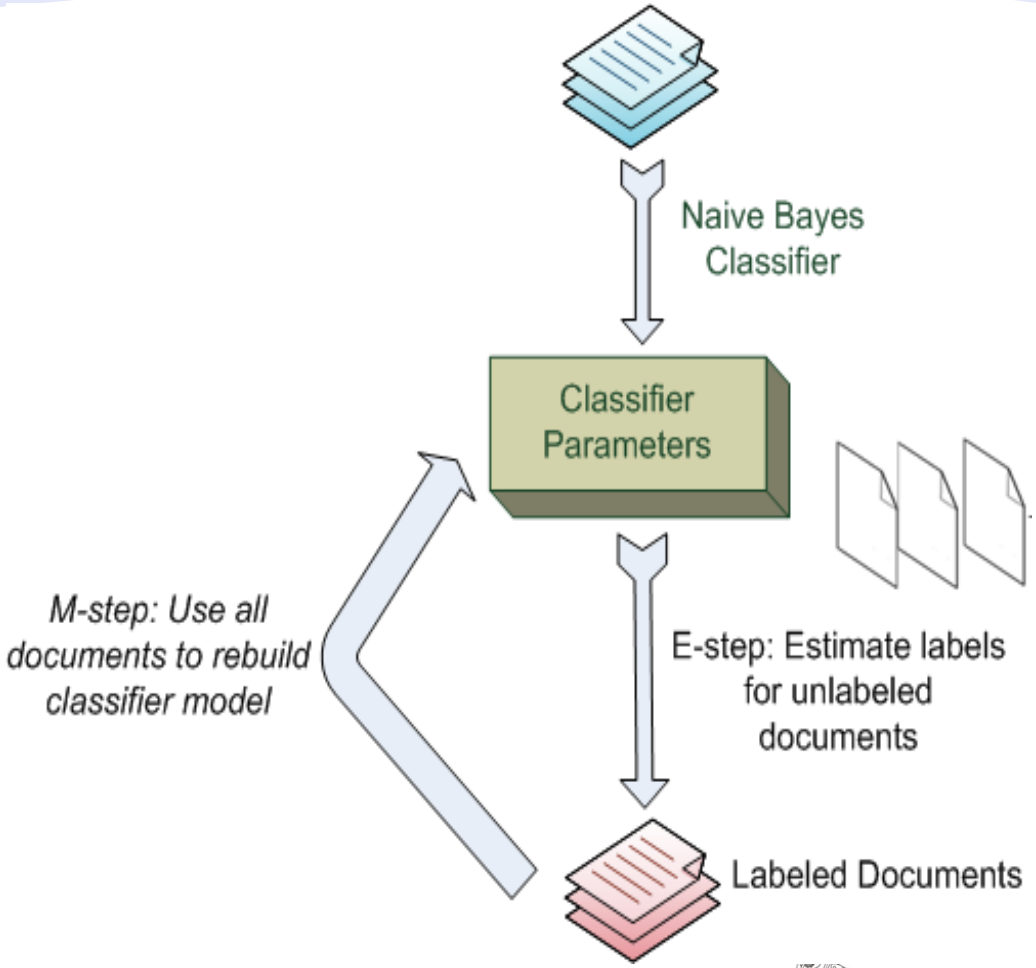
Expectation-Maximization Task

- Loop while classifier parameters improve (or number of epochs):
  - (E-step) Use the current classifier to estimate component membership of each unlabeled document $Q(h', h)$
  - (M-step) Re-estimate the classifier given the estimated component membership of each document. Use maximum a posteriori parameter estimation to find

$$h = \arg\max_{h'} Q(h'|h)$$

Categorization Algorithm

1. Build a classifier model using only labeled documents.
2. Estimate the probable labels for each unlabeled document.
3. Re-estimate the classifier using the probabilistic class labels of each document.
4. The steps 2-3 are repeated until the classifier parameters are still improved.



COPPE UFRJ

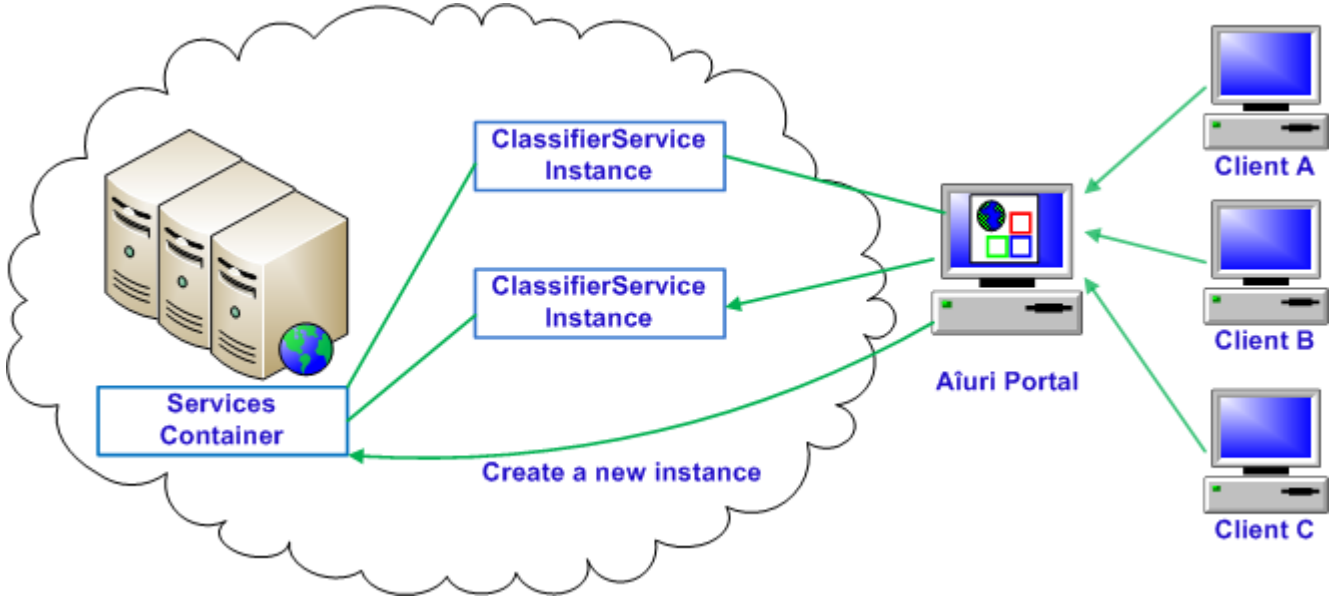
Grid Environment

- Grid NACAD
  - Globus Toolkit (GT) is an open source software toolkit used for building grids. (http://www.globus.org).
  - Globus Toolkit 4 (GT4) – version: 4.2
  - A Grid Service is a Web Service that conforms to a set of conventions (interfaces and behaviors) that define how a client interacts with a Grid Service.
  - Grid Services convert your application into a Web-application, which is published, addressable, and used through the Web.

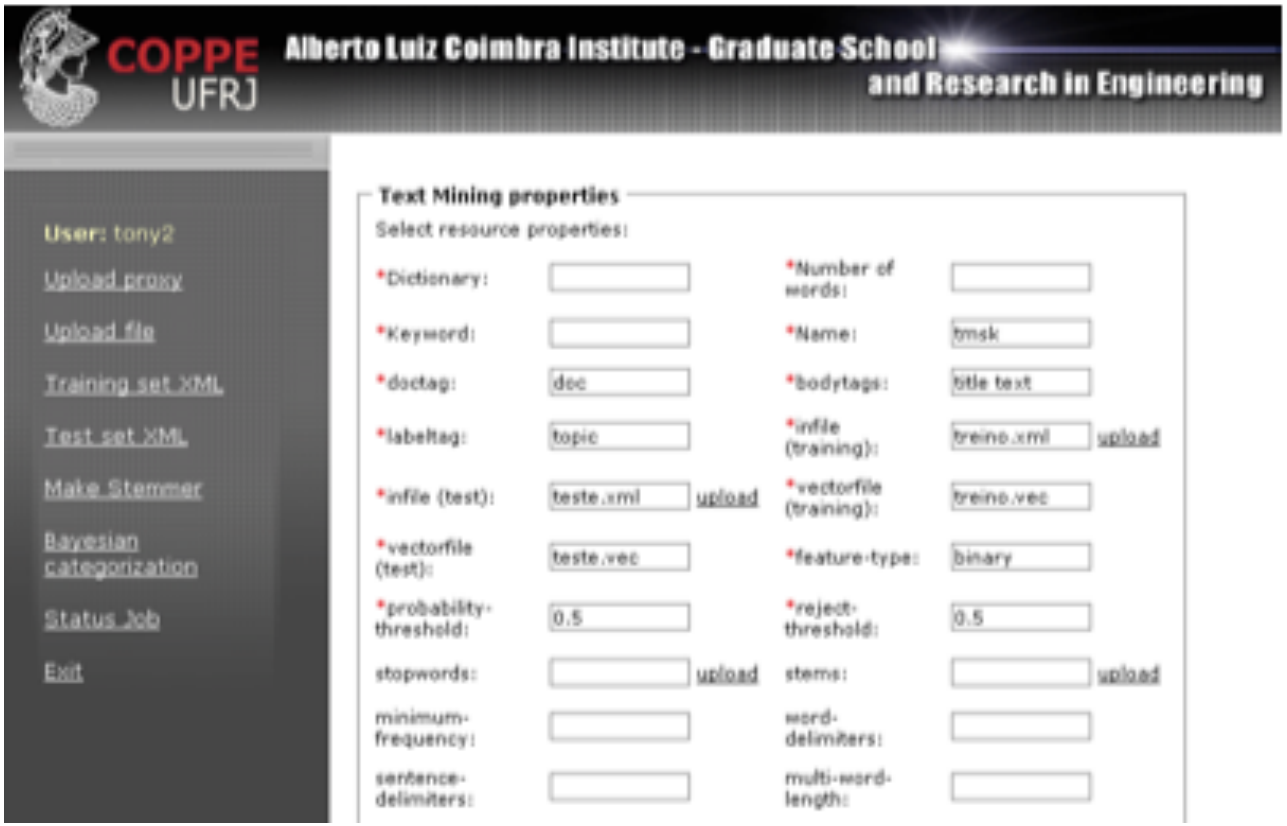
COPPE UFRJ

Aîuri Portal

- Implementation of text mining task in distributed environment allow the access of data collections and the distributed execution of tasks.
- Distribution of the classifier instances across the grid.



E-M task tunning



Testing

- Reuters-21578 Text Categorization Collection;
- The documents were assembled and indexed with categories by personnel from Reuters Ltd.;
- R8 collection - 8 of the 10 most frequent classes considering only the documents with a single topic and the classes which have at least one example for train and test;

Class	# train docs	# test docs	Total # docs
grain	41	10	51
trade	251	75	326
money-fx	206	87	293
crude	253	121	374
interest	190	81	271
ship	108	36	144
acq	1596	696	2292
earn	2840	1083	3923
Total	5485	2189	7674

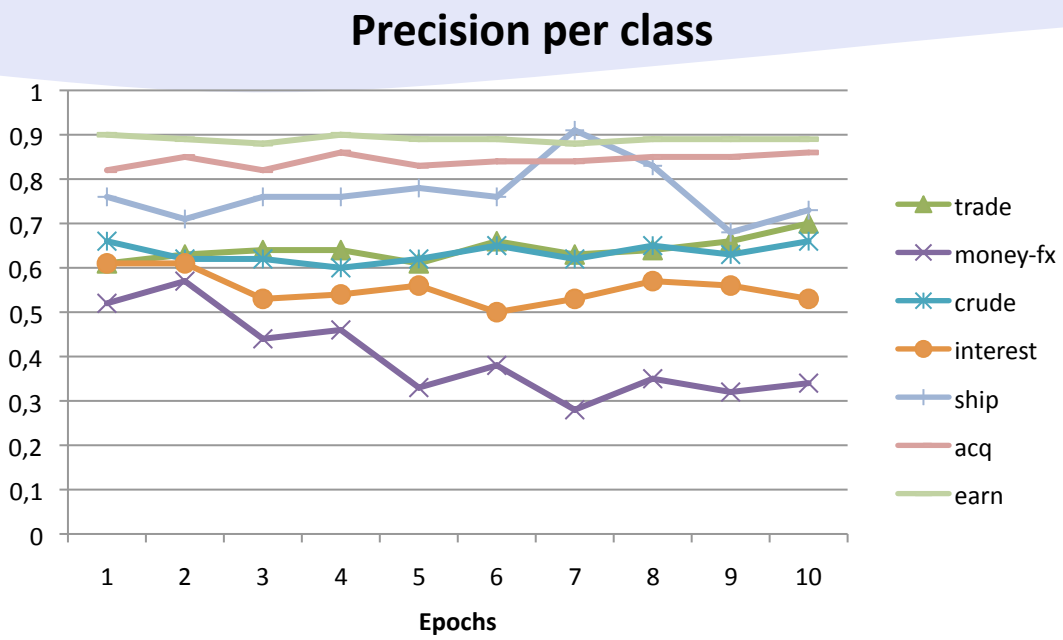
COPPE UFRJ

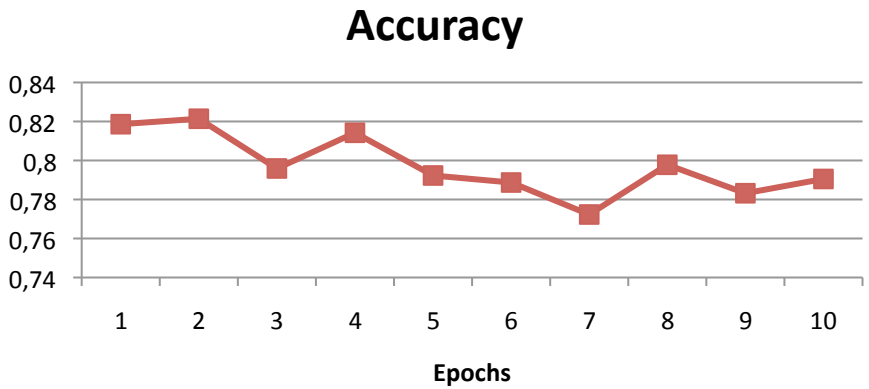
Results

- Each epoch:
  - Number of tokens in each document
  - Probabilities of each word in a category and the category itself
  - Confusion matrix of labeled documents with metrics (accuracy, precision and f-measure)
  - Confusion matrix of unlabeled documents with metrics (accuracy, precision and f-measure)

COPPE UFRJ

Results


Precision per class
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
1 2 3 4 5 6 7 8 9 10
Epochs
trade
money-fx
crude
interest
ship
acq
earn


Accuracy
0,84
0,82
0,8
0,78
0,76
0,74
1 2 3 4 5 6 7 8 9 10
Epochs

Results

confusionmatrix10.xml

	trade	money	crude	interest	ship	acq	earn
trade	24	1	0	0	0	0	0
money	1	16	0	3	0	0	1
crude	2	1	20	0	0	1	1
interest	1	10	0	8	0	0	0
ship	0	0	0	0	11	0	0
acq	4	5	1	0	1	124	24
earn	2	11	9	4	1	18	231

COPPE UFRJ

Conclusions

- The extension of the academic portal exploring distributed services across grid environments;
- Use of semi-supervised learning methods for categorization of large collection of documents;
- Extraction of consistent knowledge manipulating big sets of textual documents
- Visualization of the results and tunning of the parameters, allowing better understanding of the behavior of different knowledge discovery tasks

COPPE UFRJ

Text Classification on a Grid Environment

NACAD
High Performance Computing Center

ntt
Núcleo de Transferência de Tecnologia - UFRJ

VECPAR'10
COPPE
UFRJ