ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

TLabel: Text clustering and labelling in OLAP environment

EDA 2015
Bruxelles
2 – 3 avril 2015

TLabel

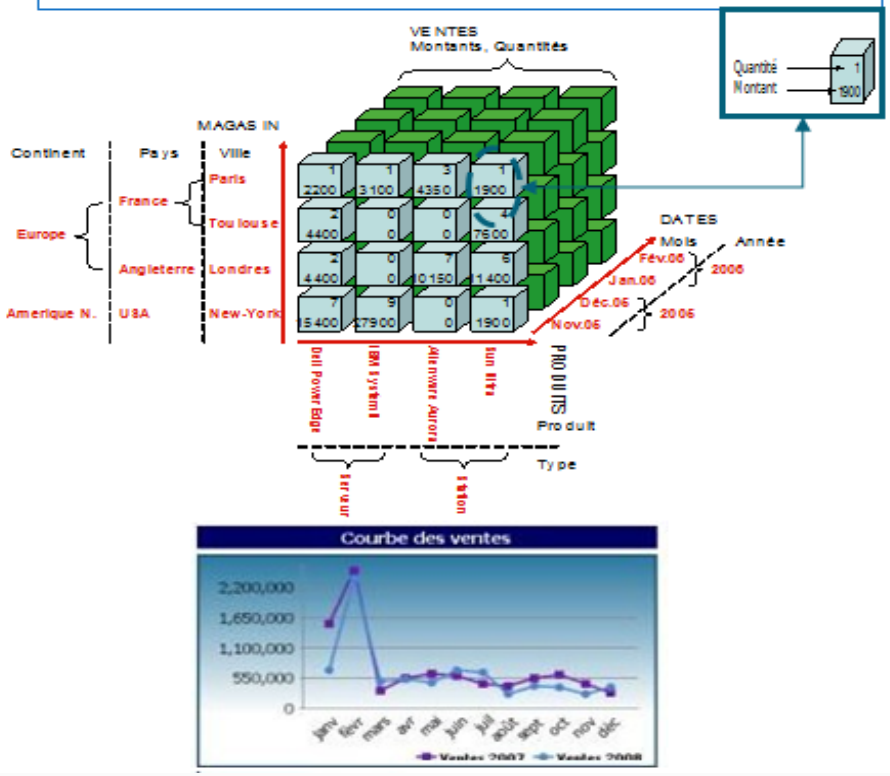
Nouvel opérateur d'agrégation
par catégorisation
dans les cubes de textes

Lamia Oukid, Omar Boussaid,
Nadjia Benblidia et Fadila Bentayeb

Context

Classical OLAP

20% structured data of information system



Text OLAP

80% non-structured data of information system

E-mails

Rapports

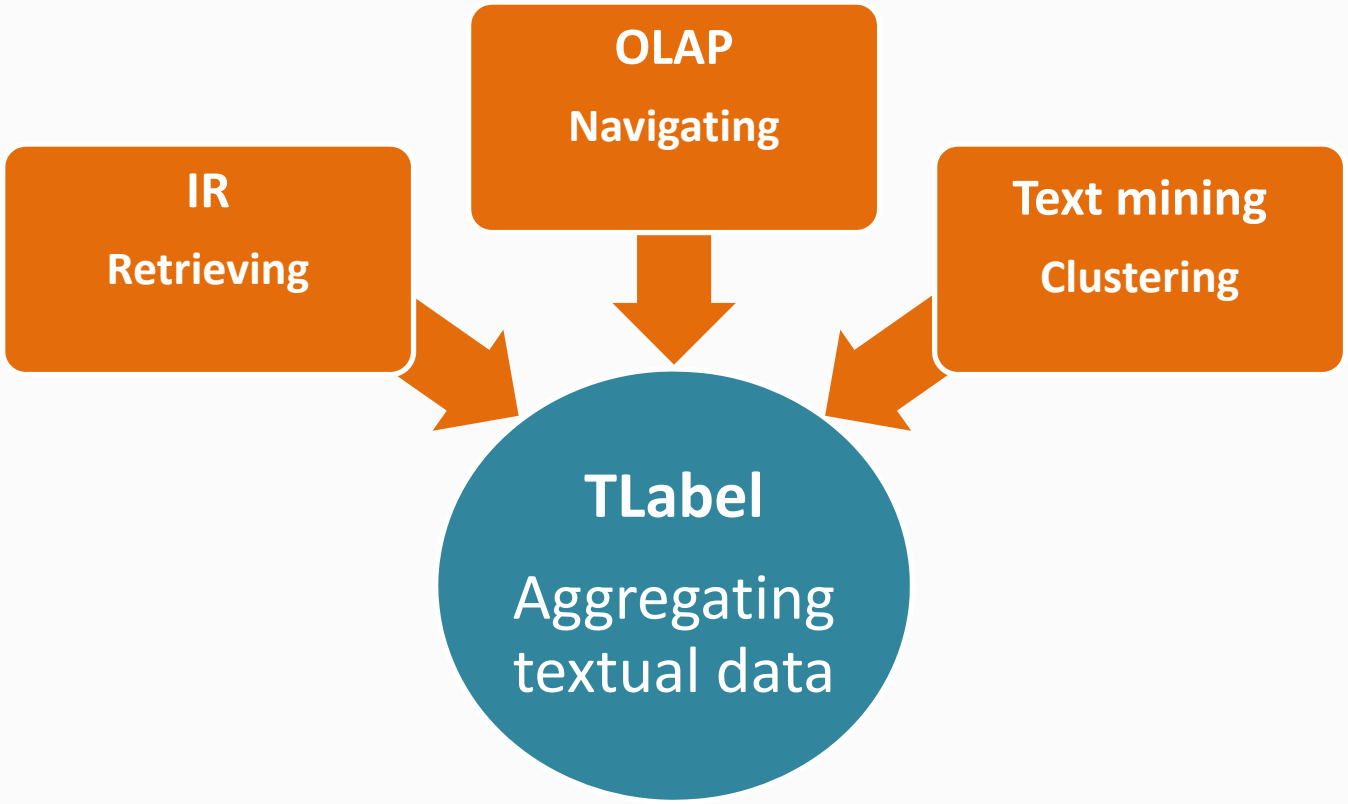
Tweets, blogs

Need new OLAP operators for text data

Towards text OLAP

- **Classical OLAP**
  - Exploring and navigating through data
- **OLAP limitation**
  - Not adapted to textual data
- **Objectives**
  - Text cubes
  - Text OLAP operators
  - Integrating data semantics in OLAP analysis

OLAP / IR/Text mining



Outline

- Text cube
- Textual query analysis
- TLabel : Clustering operator in text OLAP environment
- Experiments and results
- Conclusion and future work

Text Cube

Text definition

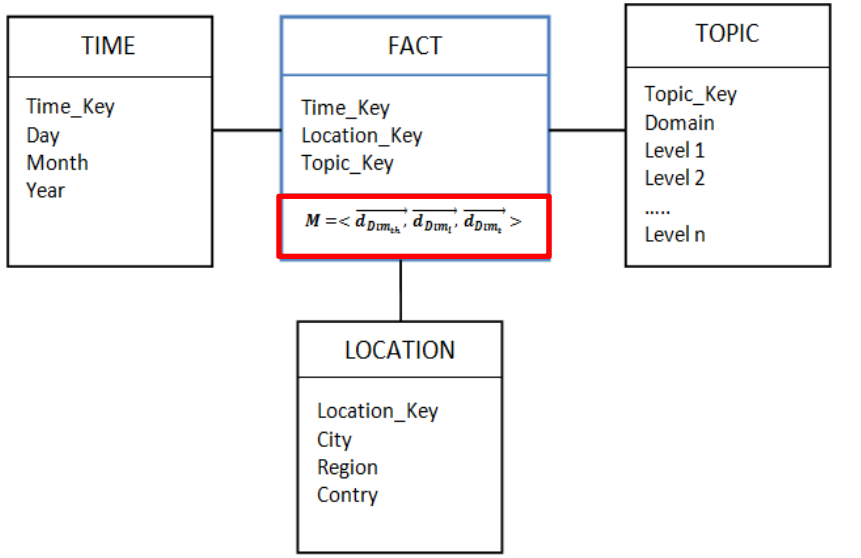
- Set of terms
  - Textuel content
- Set of metadata
  - Information on textual data
- Set of concepts
  - Extracted from domain ontology
  - Enrich text contents

Text cube modelling

- Dimensions
  - Semantic dimensions
  - Metadata dimensions
- Textual measure M
  - Vector of weighted concepts
  - One vector per dimension
- Vector of concepts

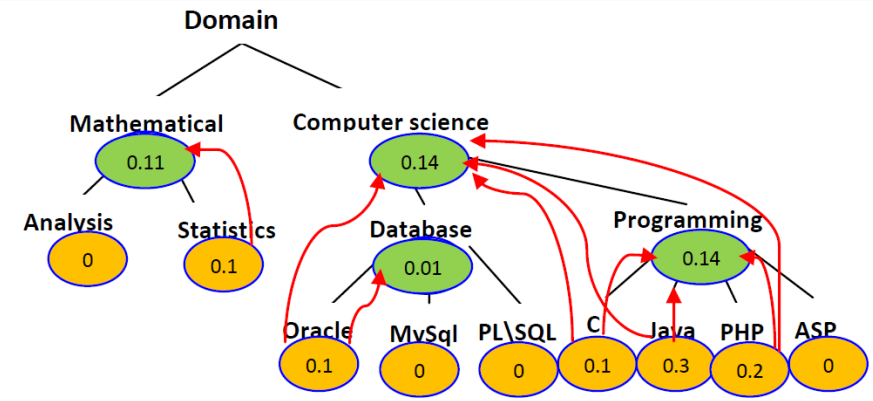
$$M = < \overrightarrow{d_{Dim_1}}, \overrightarrow{d_{Dim_2}}, \ldots, \overrightarrow{d_{Dim_*}} >$$

Text cube: Example



Text Cube Star Schema

Topic dimension : Computer science



Relevance propagation in the concept hierarchy

Text cube for CV collection

Decision Query

Query modeling

- Given text cube with n dimensions: Simple Query
  - Q = <V1, V2,…, Vn>
- Decision query with user preferences
  - The user can assign weight to each dimension
- Applied method
  - Generalized Cosinus Similarity : between query-document
- Result: Relevant text documents

Decision Query

More than extracting relevant text documents…

- Information Retrieval
  - Searching for relevant text documents
- Extracting knowledge from text documents
  - Ranking
  - Clustering
  - Resume
  - …
- Text mining
  - Supervised methods
  - No supervised methods

TLabel: Text Labelling

- Combining OLAP/IR/Text mining
- Aggregating by clustering
  - Clusters of documents
  - Adapted K-means
- Assigning labels to clusters of documents
  - Domain ontology

TLabel: Text Labelling

- Clustering step: OCluster - OLAP-Cluster
  - Adapetd K-means
  - Clusters of documents
- Labelling step
  - For each cluster, compute its DResume: resume document
  - Dresume is a vector of weighted terms
  - Assign to each cluster one label obtained from DResume

Documents Clustering

- OCluster : OLAP-Cluster
  - Input: set of documents obtained from decision query
  - Output: set of documents clusters
  - Method: K-means with similarity function OR*ank*

- *ORank*: Computes the similarity between documents

$$ORank(d, ct) = \frac{\sum_{i=1}^{n}(\alpha_i \times Sim(\overrightarrow{d_{Dim_i}}, \overrightarrow{ct_{Dim_i}}))}{n}$$

- α_i : user preferences
- n: number of dimensions

DResume

- For each cluster of documents
  - Computes its Dresume document

$$DResume = < \overrightarrow{DResume_{Dim_1}}, \overrightarrow{DResume_{Dim_2}}, ..., \overrightarrow{DResume_{Dim_*}} >$$

$$\overrightarrow{DResume_{Dim_i}} = \frac{\sum_{i=1}^{N} \overrightarrow{d_{Dim_i}}}{N}$$

N: Number of the documents in the cluster

Cluster Labelling

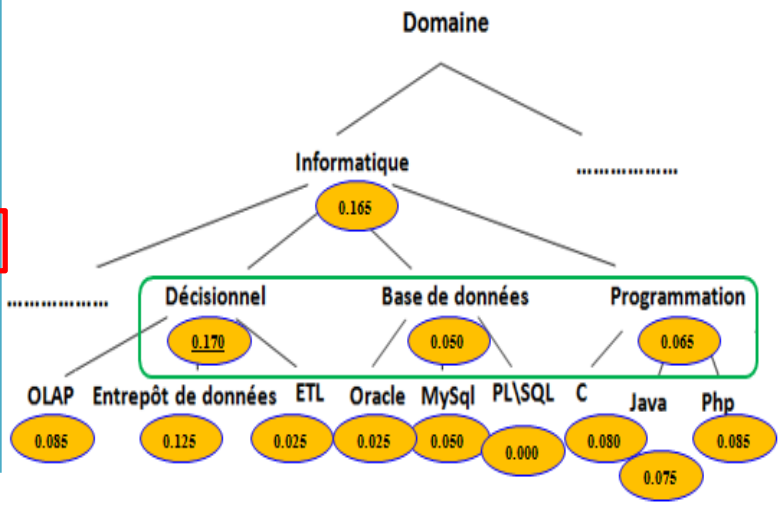
- Input
  - *DResume*
  - One dimension
  - Domain ontology
- Method
  - *Dresume* Projection on the domain ontology
- Output
  - Documents clusters labelled
  - One label for one cluster

TLabel: Example

Opérateur d'agrégation *TLabel*				THEMATIQUE	
				Domaine	Informatique
LOCALISATION	Ville	TEMPS	Année		
	France		2014		

Classe	Documents	*Label*
Cl 1	{d1, d5}	Décisionnel
Cl 2	{d3, d4}	Base de données
Cl 3	{d2}	Programmation

OLAP analysis with TLabel



Labelling documents cluster Cl 1

Experiments

- Data sources
  - 2000 CVs of candidates
  - Topic dimension : Computer science
- Ontology: hierarchy of concepts
  - Wikipedia

Experiments & results

Experiments

- Preparing data sources
  - Text Tokenisation
  - Drop stop words
  - Term Lemmatisation: Tree tagger
- Loading data into text cube
  - Semantic dimensions: Topic and Location
  - For each dimension, load the concept hierarchy from the corresponding domain ontology
  - Time Dimension

Experiments & results

Results

- Query: <Topic= Computer Science, Location= France , Time= 2014>

- ***Ocluster***

OCluster	Cl 1	Cl 2	Cl 3	Cl 4	Cl 5	Cl 6	Cl 7	Cl 8
Documents Number	98	181	84	179	215	178	1	216
Total	1152							

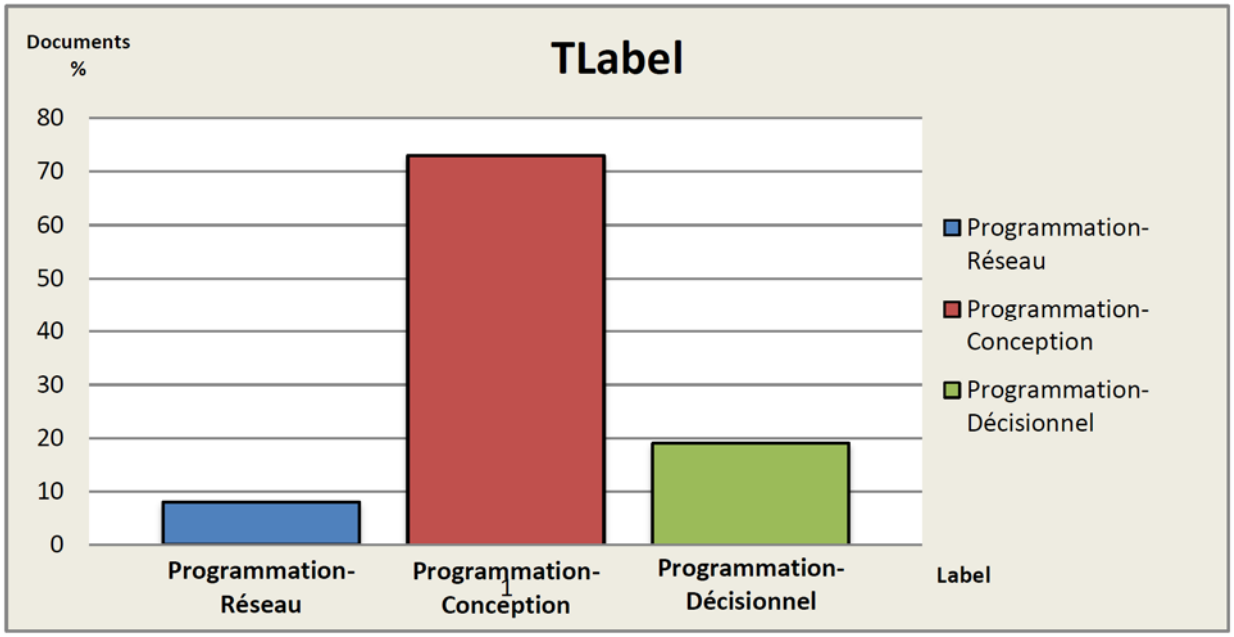
- **Labelling**

TLabel	Programmation-Réseau	Programmation-Conception	Programmation-Décisionnel
OCluster	Cl 1, Cl 7	Cl 2, Cl 3, Cl 4, Cl 6, Cl 8	Cl 5
Documents Number	99 (8%)	838 (73 %)	215 (19%)

Experiments & results

Results

- Query: <Topic= Computer Science, Location= France, Time= 2014 >

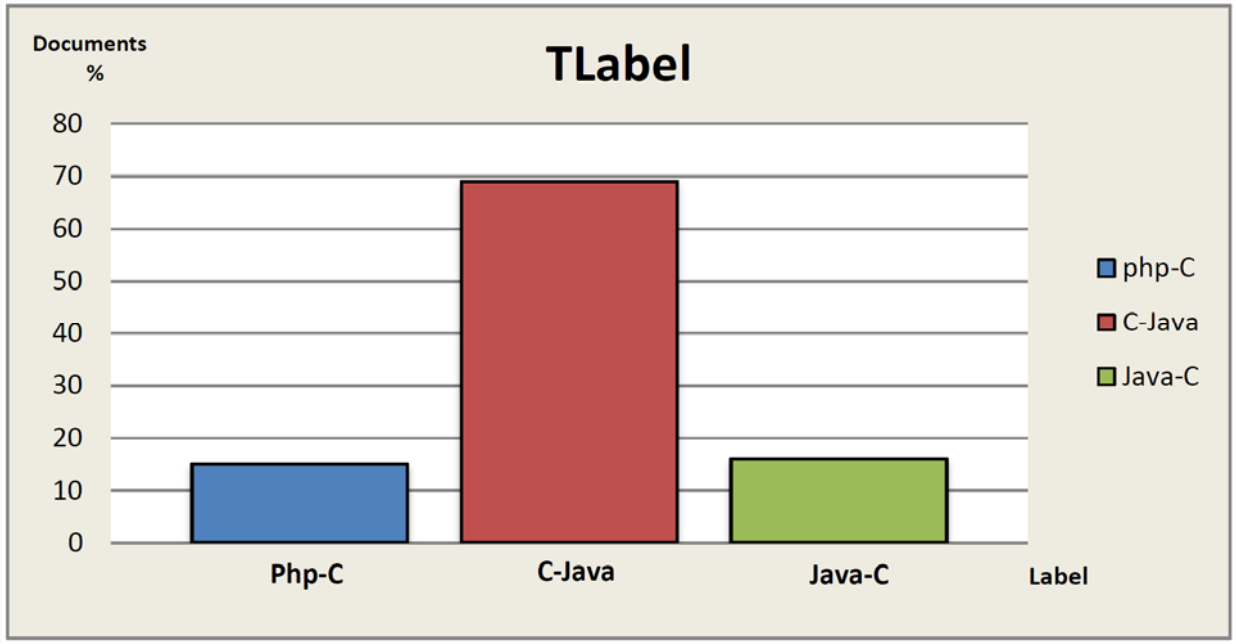


OLAP with TLabel

Experiments & results

Results

- Drill-down on Topic dimension



Drill-Down on Topic dimension with TLabel

Conclusion & Future Work

Conclusion

- TextLabel : Clustering text documents in text OLAP systems
- Text mining
  - *Ocluster*: adapting K-means in OLAP environment
- Documents clusters labelling
  - Dresume
  - Domain ontology
- Experiments on CV collections

Conclusion & perspectives

Future Work

- Think about other methods to obtain *Dresume document*
- Cluster Labelling according to several dimensions
- Evaluating TLabel with other text collections
- Validating TLbel with known labelled clusters