TLabel: Text clustering and labelling in OLAP environment

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

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

OLAP Navigating

IR Retrieving

Text mining Clustering

TLabel
Aggregating textual data
Outline

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

Text cube modelling

- Dimensions
  - Semantic dimensions
  - Metadata dimensions

- Textual measure $M$
  - Vector of weighted concepts
  - One vector per dimension

- Vector of concepts

$$M = < d_{Dim_1}, d_{Dim_2}, ..., d_{Dim_*} >$$
Text Cube

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
  - $Q = <V_1, V_2, ..., V_n>$

- 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
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
TLlabel: Text Labelling

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

- **Clustering step**: OCluster - OLAP-Cluster
  - Adapeted 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 ORank

- **ORank**: Computes the similarity between documents
  \[
  ORank(d, ct) = \sum_{i=1}^{n} (\alpha_i \times Sim(d_{Dim_i}, ct_{Dim_i})) / n
  \]
  - \(\alpha_i\): user preferences
  - \(n\): number of dimensions
For each cluster of documents

- Computes its DResume document

\[
\text{DResume} = \langle \overrightarrow{DResume_{Dim_1}}, \overrightarrow{DResume_{Dim_2}}, \ldots, \overrightarrow{DResume_{Dim_n}} \rangle
\]

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

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
OLAP analysis with TLabel

Labelling documents cluster Cl 1
Experiments & results

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
## Results

**Query:** \(<\text{Topic= Computer Science, Location= France, Time= 2014}>\)

### OCluster

<table>
<thead>
<tr>
<th>OCluster</th>
<th>Cl 1</th>
<th>Cl 2</th>
<th>Cl 3</th>
<th>Cl 4</th>
<th>Cl 5</th>
<th>Cl 6</th>
<th>Cl 7</th>
<th>Cl 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents Number</td>
<td>98</td>
<td>181</td>
<td>84</td>
<td>179</td>
<td>215</td>
<td>178</td>
<td>1</td>
<td>216</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1152</td>
</tr>
</tbody>
</table>

### Labelling

<table>
<thead>
<tr>
<th>TLabel</th>
<th>Programmation-Réseau</th>
<th>Programmation-Conception</th>
<th>Programmation-Décisionnel</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCluster</td>
<td>Cl 1, Cl 7</td>
<td>Cl 2, Cl 3, Cl 4, Cl 6, Cl 8</td>
<td>Cl 5</td>
</tr>
<tr>
<td>Documents Number</td>
<td>99 (8%)</td>
<td>838 (73 %)</td>
<td>215 (19%)</td>
</tr>
</tbody>
</table>
Results

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

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