Cube de textes et opérateur d'agrégation basé sur un modèle vectoriel adapté

Text Cube Model and aggregation operator based on an adapted vector space model

Lamia Oukid, Ounas Asfari, Fadila Bentayeb, Nadjia Benblidia, Omar Boussaid

University of Blida, Saad Dahlab (LRDSI Laboratory)
University of Lyon (ERIC Laboratory, Lyon 2)

Presentation: Lamia Oukid

14 Juin 2013

Outline

- OLAP analysis for textual data !

- OLAP analysis for textual data !

1. How can we represent and analyze textual data ?

- OLAP analysis for textual data !

1. How can we represent and analyze textual data ?
2. How can we aggregate textual data ?

- Requirements

1. Propose aggregation operators for textual data
2. Extend the classical star schema model

IR VS. OLAP

Information Retrieval IR

- User queries of keywords
- List of relevant documents
  - keywords' frequency in the document
- Extracting information from documents

IR VS. OLAP

Information Retrieval IR

- User queries of keywords
- List of relevant documents
  - keywords' frequency in the document
- Extracting information from documents

On-Line Analytical Processing OLAP

- Analysis queries
- Results relevant to the decision-making
- *Roll up*, *Drill down* from one view to another

IR VS. OLAP

Information Retrieval IR

- User queries of keywords
- List of relevant documents
  - keywords' frequency in the document
- Extracting information from documents

On-Line Analytical Processing OLAP

- Analysis queries
- Results relevant to the decision-making
- *Roll up*, *Drill down* from one view to another

Combination IR / OLAP

- Extract relevant information from documents to define measures for textual data analysis
- Define OLAP operator adapted to textual data

Contributions

1. *TCube* : Text Cube model
   - Semantics dimensions –> domain ontology
   - Textual analysis measure –> relevance propagation
2. *ORank* : Textual Aggregation operator (OLAP-Rank)
   - Adaptation of the vector space model to OLAP analysis
   - Integration of decision-maker's preferences

Related work

Extend the classical data cube model

- Zhang et al.(2009) *Topic Cube*
- Pérez et al.(2007) *R-Cube*
- Lin et al.(2008) *Text Cube*
- Zhang et al.(2011) *MitexCube*

Related work

Extend the classical data cube model

- Zhang et al.(2009) *Topic Cube*
- Pérez et al.(2007) *R-Cube*
- Lin et al.(2008) *Text Cube*
- Zhang et al.(2011) *MitexCube*

Aggregation operator for textual data

- Ravat et al.(2007) *AVG-KW*
- Ravat et al.(2008) *TOP-KWk*
- Ben-Messaoud et al.(2004) *OpAC*
- Bringay et al.(2011) *Tf-Idf adaptative*

TCube

- $Cube = (F, Dim_1, Dim_2, ..., Dim_*, M_1, M_2, ..., M_*)$

TCube

- $Cube = (F, Dim_1, Dim_2, ..., Dim_*, M_1, M_2, ..., M_*)$
- *Semantic dimension*
  - $Dim_r, r \in [1, *]$ include attributes $A =< a_1, a_2, ..., a_* >$ organized in hierarchical levels $< l_1, l_2, ..., l_* >$
  - $l_i =< c_1, c_2, ...c_n >$ includes a set of concepts $c_j, (j \in [1, n])$ from domain ontology

TCube

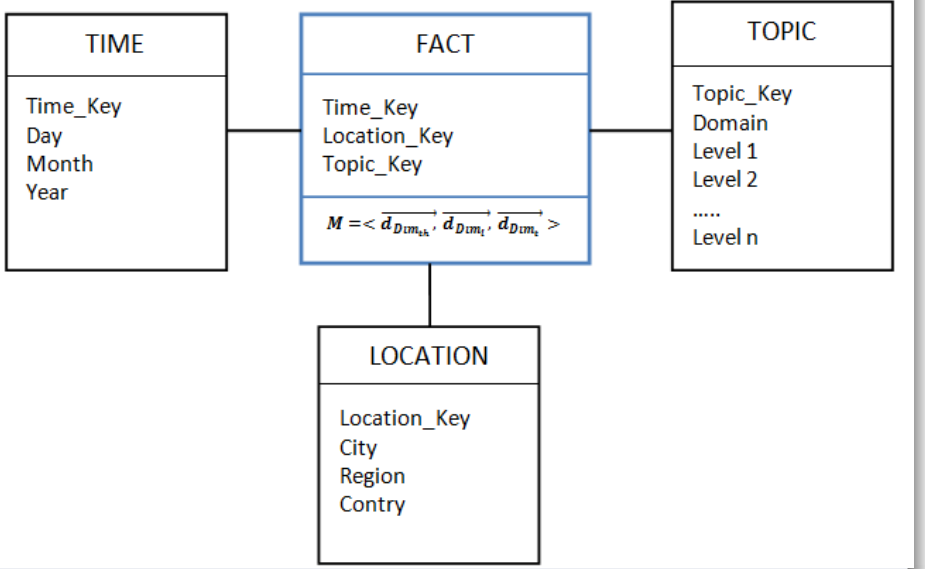
- $Cube = (F, Dim_1, Dim_2, ..., Dim_*, M_1, M_2, ..., M_*)$
- *Semantic dimension*
  - $Dim_r, r \in [1, *]$ include attributes $A =< a_1, a_2, ..., a_* >$ organized in hierarchical levels $< l_1, l_2, ..., l_* >$
  - $l_i =< c_1, c_2, ...c_n >$ includes a set of concepts $c_j, (j \in [1, n])$ from domain ontology
- Textual analysis measure *M*
  - *M* represents each document *d* by several weighted concepts vectors, one vector for each dimension Dim_r of *Cube*

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

$$\overrightarrow{d_{Dim_r}} =< w_{c_1}, w_{c_2}, ..., w_{c_n} >$$

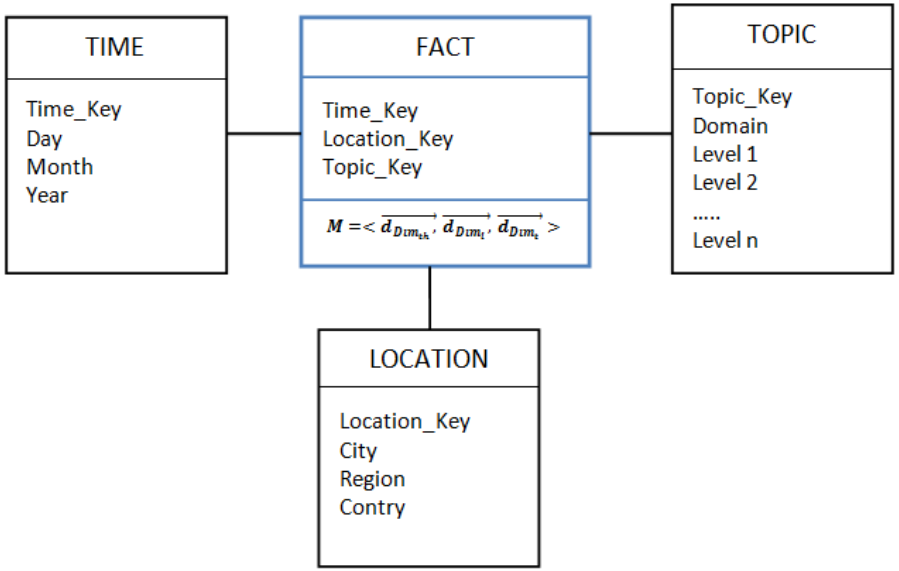
OLAP analysis on CVs' collection

Star schema of *TCube*

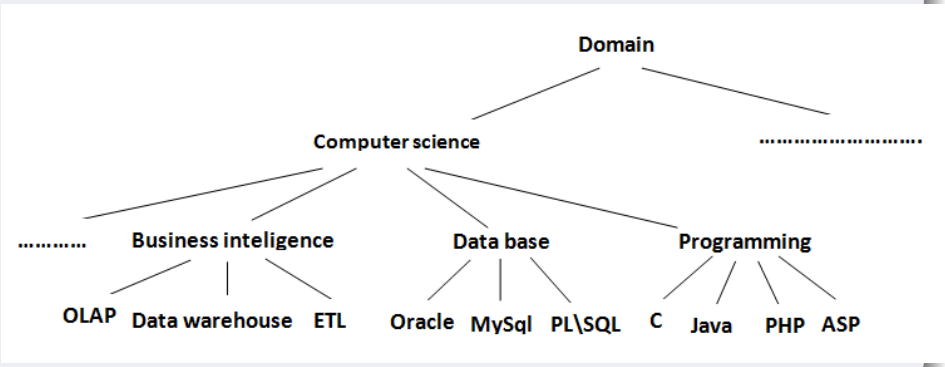


OLAP analysis on CVs' collection

Star schema of *TCube*



Example of a topics' hierarchy



Relevance propagation

- Consider text semantics during the computation of the term weight in a document

Relevance propagation

- Consider text semantics during the computation of the term weight in a document
- Reassign scores of concepts in the hierarchy

Relevance propagation

1. Compute the weights of the document terms existing in the concepts hierarchy (*Term Frequency Tf*)

$$Tf_{t,d} = \frac{n_{t,d}}{N_d} \quad (1)$$

$n_{t,d}$: occurrence frequency of term t in document d

N_d : total terms number of the document d

Relevance propagation

1. Compute the weights of the document terms existing in the concepts hierarchy (*Term Frequency Tf*)

$$Tf_{t,d} = \frac{n_{t,d}}{N_d} \quad (1)$$

$n_{t,d}$: occurrence frequency of term t in document d

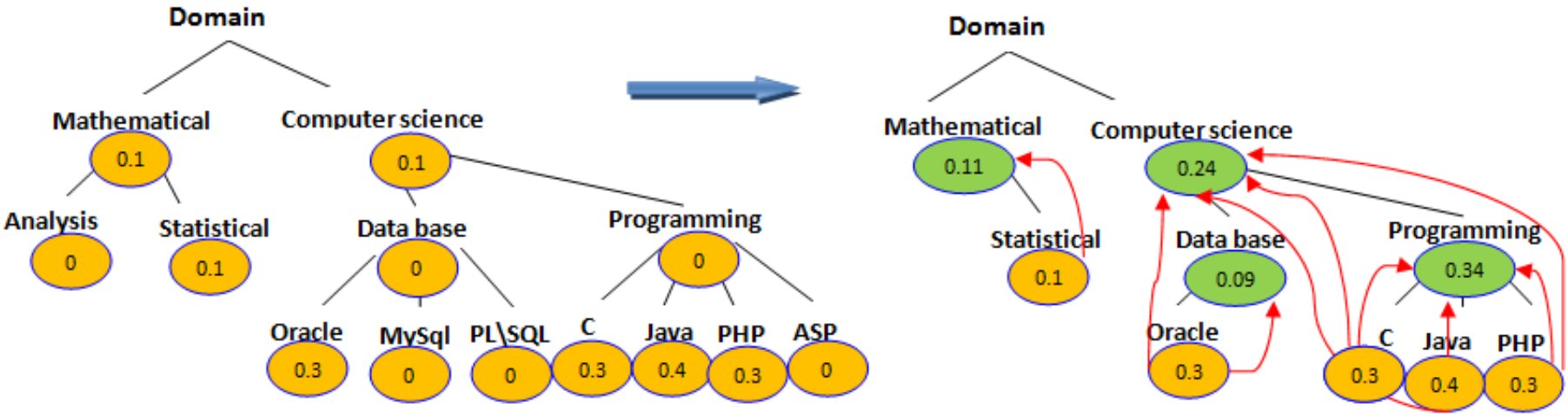
N_d : total terms number of the document d

2. Compute weights of ancestors of each leaf node which has a weight not null

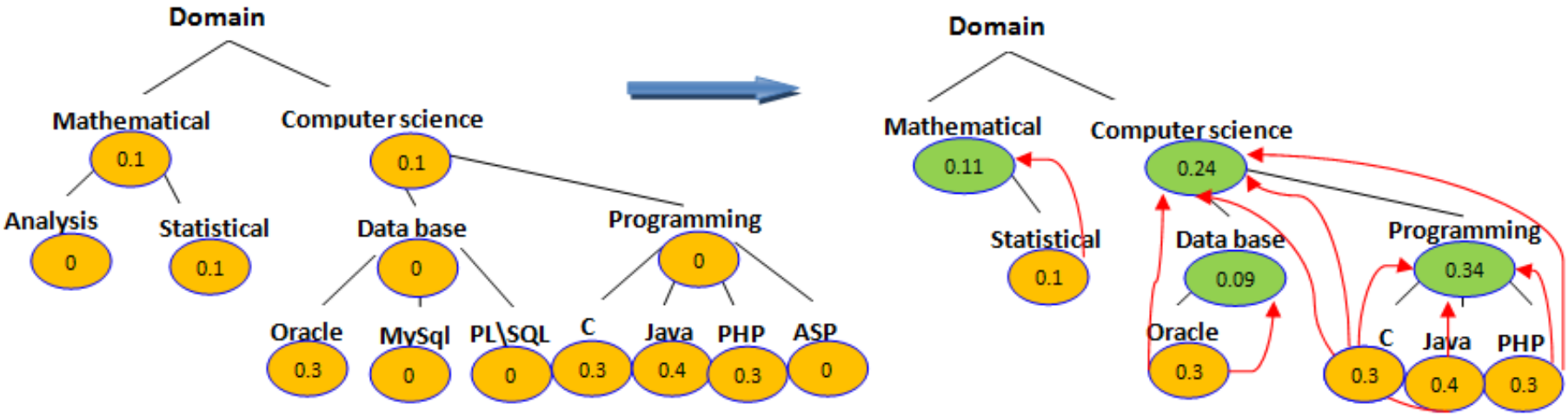
$$Poids(n_k, n_{fi}) = Poids(n_k) + Poids(n_{fi})^{distance(n_k, n_{fi})+1} \quad (2)$$

n_{fi} : leaf node from which the relevance propagation is performed
n_k : ancestor node to which the new weight is computed
$distance(n_k, n_{fi})$: semantic distance between the nodes n_k and n_{fi}

Relevance propagation for the TOPIC dimension (Example)



Relevance propagation for the TOPIC dimension (Example)



Result

$\overrightarrow{d_{Dim_{th}}} =<$
computer-science(0.24), *programming*(0.34), *java*(0.4), *php*(0.3), *c*(0.3), *database*(0.09), *oracle*(0.3),

mathematical(0.11), *statistics*(0.1) $>$

ORank operator

- Aggregate the semantics content of the documents by ranking them

ORank operator

- Aggregate the semantics content of the documents by ranking them
- Combine the different vector spaces of *TCube* dimensions

ORank operator

- Analysis query $Q =< \overrightarrow{q_1}, \overrightarrow{q_2}, ..., \overrightarrow{q_*} >$ where q_r is the query at the dimension Dim_r
- The aggregate of a document $d =< \overrightarrow{d_{Dim_1}}, \overrightarrow{d_{Dim_2}}, ..., \overrightarrow{d_{Dim_*}} >$ is calculated as follows :

$$ORank(d) = \sum_{i=1}^{n} (\alpha_i \times Sim(\overrightarrow{d_{Dim_i}}, \overrightarrow{q_i})) \quad (3)$$

$\overrightarrow{d_{Dim_i}}$: document vector at the dimension Dim_i

$Sim(\overrightarrow{d_{Dim_i}}, \overrightarrow{q_i})$: cosine similarity between $\overrightarrow{d_{Dim_i}}$ and $\overrightarrow{q_i}$

Decision-maker's preferences in *ORank*

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

- α_i –> user can indicate the importance of each *TCube* dimension as percentage P_i

$$\alpha_i = P_i \times n \quad (4)$$

- Documents are ranked in descending order of $ORank(d)$

Example of analysis on a *TCube*

Aggregation operator *ORank*				TOPIC						
				Domain	Computer science					
				level 1	Business intelligence		Database		Programming	
LOCATION	Country	TIME	Year		Doc	*Rank*	Doc	*Rank*	Doc	*Rank*
	France		2010		CV 1	1	CV 1	2	CV 1	4
					CV 2	4	CV 2	4	CV 2	1
					CV 3	3	CV 3	1	CV 3	3
					CV 4	2	CV 4	3	CV 4	2

(a)

Roll-up on TOPIC dimension

Aggregation operator *ORank*				TOPIC		
				Domain	Computer science	
LOCATION	Country	TIME	Year		Doc	*Rank*
	France		2010		CV 1	1
					CV 2	3
					CV 3	2
					CV 4	4

(b)

Data set

- Collection of 1000 documents (CVs)
  - Candidates for a Master 2 in Business Intelligence
- Construct concepts' hierarchies
  - Thematic portal from Wikipedia
  - Geographical ontology Geonames

1. Pretreatment

- *Text Tokenisation*, text conversion to a lowercase, stop-words removal, *terms lemmatisation* (*tree tagger*).

1. Pretreatment

- *Text Tokenisation*, text conversion to a lowercase, stop-words removal, *terms lemmatisation* (*tree tagger*).

2. Feeding *TCube* dimensions

- Develop concepts hierarchy as XML tree for each semantic dimension

1. Pretreatment

- *Text Tokenisation*, text conversion to a lowercase, stop-words removal, *terms lemmatisation* (*tree tagger*).

2. Feeding *TCube* dimensions

- Develop concepts hierarchy as XML tree for each semantic dimension

3. Tests

- Analysis measure : comparison between *Tf* measure and our *text analysis measure*
- *ORank* textual aggregation operator : analysis queries

Comparison between *Tf* measure and our *text analysis measure*

Tf measure	Our *text analysis measure M*
M < $\overrightarrow{d}$ = donnée (0.0348), base (0.0348), statistique (0.0290), informatique (0.0232), lyon (0.0232), analyse (0.0174), programmation (0.0174), compétence (0.0116), mention (0.0116), professionnel (0.0116), système (0.0116), lumière (0.0116). >	M < $\overrightarrow{d_{Dim_{th}}}$ = informatique (0.0239), statistique (0.0293), programmation (0.0180), réseau (0.0119), décisionnel (0.0065), php (0.0059), java (0.0059), r (0.0059), sql (0.0059), access (0.00598), base donnée (3.5856E-5), bureautique (3.5856E-5). $\overrightarrow{d_{Dim_l}}$ = lyon (0.0232), rhône-alpes (0.0005), france (0.0125E-3) $\overrightarrow{d_{Dim_t}}$ = 13 Avril 2010 (1.0) >

Comparison between *Tf* measure and our *text analysis measure*

Tf measure	Our *text analysis measure M*
M < $\overrightarrow{d}$ = donnée (0.0348), base (0.0348), statistique (0.0290), informatique (0.0232), lyon (0.0232), analyse (0.0174), programmation (0.0174), compétence (0.0116), mention (0.0116), professionnel (0.0116), système (0.0116), lumière (0.0116). >	M < $\overrightarrow{d_{Dim_{th}}}$ = informatique (0.0239), statistique (0.0293), programmation (0.0180), réseau (0.0119), décisionnel (0.0065), php (0.0059), java (0.0059), r (0.0059), sql (0.0059), access (0.00598), base donnée (3.5856E-5), bureautique (3.5856E-5). $\overrightarrow{d_{Dim_l}}$ = lyon (0.0232), rhône-alpes (0.0005), france (0.0125E-3) $\overrightarrow{d_{Dim_t}}$ = 13 Avril 2010 (1.0) >

Applying *ORank* on training Data

- Query 1 : TIME = 2010, LOCATION = "Lyon" et TOPIC= "Informatique décisionnelle"
- Query 2 : TIME = 2010, LOCATION = "Lyon" et TOPIC= "Informatique décisionnelle", preferences : 50% TOPIC et 25% TIME and LOCATION.

Applying *ORank* on training Data

Results of *ORank* on query 1

$ORank(d)$	Documents	$Sim(d_{Dim_{th}}, q_{th})$	$Sim(d_{Dim_{l}}, q_{l})$	$Sim(d_{Dim_{t}}, q_{t})$
1(2.6314)	CV 8	0.6314	1.0	1.0
2(2.5070)	CV 4	0.5070	1.0	1.0
3(2.4901)	CV 1	0.4901	1.0	1.0
4(2.4856)	CV 10	0.4856	1.0	1.0
5(2.4761)	CV 7	0.5477	0.9284	1.0
6(2.4648)	CV 9	0.4648	1.0	1.0
7(2.3281)	CV 6	0.3281	1.0	1.0
8(2.2817)	CV 2	0.5746	0.7071	1.0
9(2.2778)	CV 3	0.5707	0.7071	1.0
10(2.2734)	CV 5	0.3790	0.8944	1.0

ORank results

Query 1

$ORank(d)$	Documents	$Sim(d_{Dim_{th}}, q_{th})$	$Sim(d_{Dim_{l}}, q_{l})$	$Sim(d_{Dim_{t}}, q_{t})$
1(2.6314)	CV 8	0.6314	1.0	1.0
2(2.5070)	CV 4	0.5070	1.0	1.0
3(2.4901)	CV 1	0.4901	1.0	1.0
4(2.4856)	CV 10	0.4856	1.0	1.0
5(2.4761)	CV 7	0.5477	0.9284	1.0
6(2.4648)	CV 9	0.4648	1.0	1.0
7(2.3281)	CV 6	0.3281	1.0	1.0
8(2.2817)	CV 2	0.5746	0.7071	1.0
9(2.2778)	CV 3	0.5707	0.7071	1.0
10(2.2734)	CV 5	0.3790	0.8944	1.0

ORank results

Query 1

$ORank(d)$	Documents	$Sim(d_{Dim_{th}}, q_{th})$	$Sim(d_{Dim_l}, q_l)$	$Sim(d_{Dim_t}, q_t)$
1(2.6314)	CV 8	0.6314	1.0	1.0
2(2.5070)	CV 4	0.5070	1.0	1.0
3(2.4901)	CV 1	0.4901	1.0	1.0
4(2.4856)	CV 10	0.4856	1.0	1.0
5(2.4761)	CV 7	0.5477	0.9284	1.0
6(2.4648)	CV 9	0.4648	1.0	1.0
7(2.3281)	CV 6	0.3281	1.0	1.0
8(2.2817)	CV 2	0.5746	0.7071	1.0
9(2.2778)	CV 3	0.5707	0.7071	1.0
10(2.2734)	CV 5	0.3790	0.8944	1.0

Query 2

$ORank(d)$	Documents	$Sim(d_{Dim_{th}}, q_{th})$	$Sim(d_{Dim_l}, q_l)$	$Sim(d_{Dim_t}, q_t)$
1(2.4471)	CV 8	0.6314	1.0	1.0
2(2.2678)	CV 7	0.5477	0.9284	1.0
3(2.2605)	CV 4	0.5070	1.0	1.0
4(2.2351)	CV 1	0.4901	1.0	1.0
5(2.2284)	CV 10	0.4856	1.0	1.0
6(2.1972)	CV 9	0.4648	1.0	1.0
7(2.1422)	CV 2	0.5746	0.7071	1.0
8(2.1363)	CV 3	0.5707	0.7071	1.0
9(1.9921)	CV 6	0.3281	1.0	1.0
10(1.9893)	CV 5	0.3790	0.8944	1.0

Conclusion

- New approach of OLAP analysis for textual data

Conclusion

- New approach of OLAP analysis for textual data
  - *TCube* : Text Cube Model
    - Coupling IR and OLAP
    - Text analysis Measure
    - *Semantic dimensions*

Conclusion

- New approach of OLAP analysis for textual data
  - *TCube* : Text Cube Model
    - Coupling IR and OLAP
    - Text analysis Measure
    - *Semantic dimensions*
  - *ORank* : New text aggregation operator
    - Aggregates the documents by ranking
    - Includes decision-maker's preferences

Conclusion

- Experimental study on CV's collection
  - Results show the importance of our approach for :
    - Textual data analysis
    - Textual data integration into a data warehouse
    - Improve the decision-making

Perspectives

- Measure the quality of the terms extracted by our text analysis measure
- Improve and evaluate the results of *ORank* operator

Perspectives

- Measure the quality of the terms extracted by our text analysis measure
- Improve and evaluate the results of *ORank* operator
- Evaluate the scalability of our approach

Thank you for your attention