



Top_Keyword: Agrégation de mots-clefs dans un environnement OLAP



Franck Ravat, Olivier Teste,
Ronan Tournier, Gilles Zurfluh.
**IRIT: Institut de Recherche en Informatique
de Toulouse.**
tournier@irit.fr

Plan



■ Introduction : contexte

- Contexte : Analyse en ligne (OLAP)
- Agrégation : ex. des opérations de forage
- Analyse en ligne de données textuelles

■ Modèle conceptuel

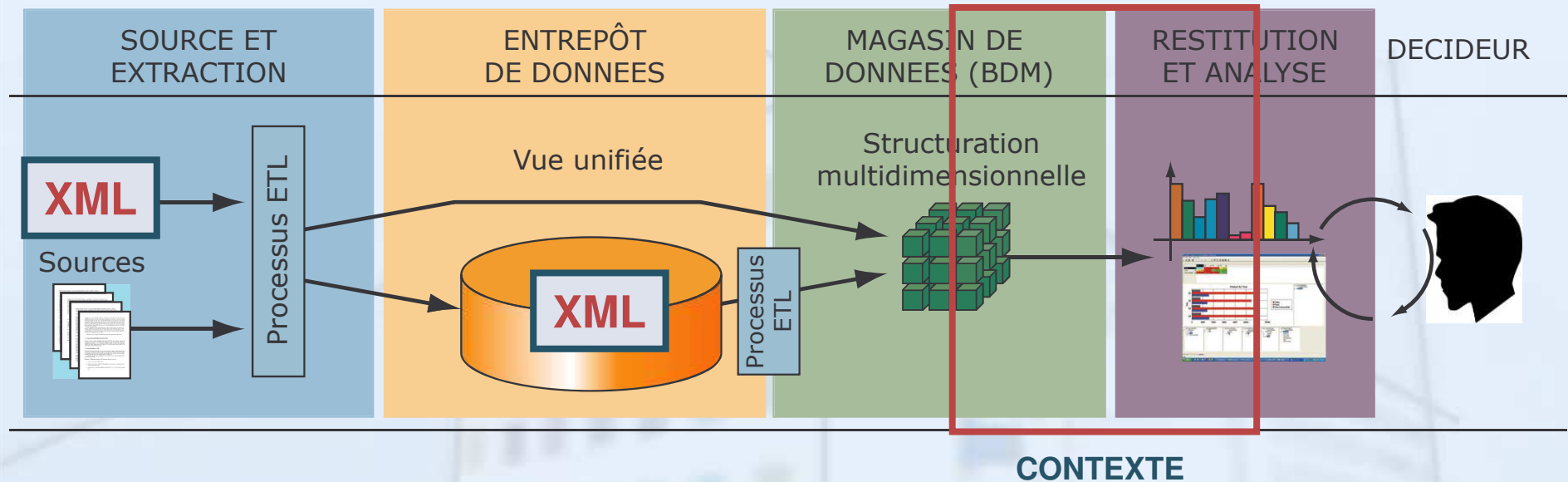
■ Fonction d'agrégation

■ Conclusion





■ Processus d'analyse en ligne



Agrégation



■ OLAP: agrégation, ex. opération de forage

COUNT (NB_Keywords)		TIME					
		Year	2005			2006	
		Month	Sept.	Nov.	Dec.	Jan.	March
AUTHOR	IdA						
	Au1	3	4	2	2	2	
	Au2	2	2	3	6	7	

COUNT (NB_Keywords)		TIME	
		Year	2005
AUTHOR	IdA		
	Au1	9	4
	Au2	7	13

Processus d'agrégation :
un *comptage* est calculé
par **Mois** puis par **Année**

- Roll-Up: Time.Month => Time.Year
- Opération inverse : Drill-Down



- Pourquoi se soucier des documents ?
 - Analyse classique : données transactionnelles
 - Données transactionnelles : 20% des données d'un système d'information [Sullivan-01], [Tseng-06]
 - **80% des données hors de portée d'OLAP**
- Limite :
 - Systèmes OLAP : indicateurs numériques
 - Texte = données non structurées à première vue
- XML : une solution pour intégrer texte et OLAP
 - Format semi-structuré
 - Disponibilité croissante de sources
 - Permet une structuration du contenu de documents

■ Documents XML

■ Plusieurs travaux

- Analyse de contenu par mots-clefs [McCabe-00]..
- Agrégation de données XML [Wang-03]...

■ Basé sur des indicateurs numériques

■ OLAP, XML et données textuelles

- Environnement d'analyse adapté [Park-05]
(peu de détails)

■ Modèle multidimensionnel

- Pas adapté a l'analyse d'un contenu textuel
- Pas d'indicateurs textuels

■ Agrégation de données textuelles

- Texte : données systématiquement non-additives !
- Agrégation : fonctions génériques COUNT et LIST

■ Systèmes OLAP : agrégation numérique

Comment effectuer une analyse avec des indicateurs textuels ?



■ Introduction : contexte

■ **Modèle conceptuel**

- Concepts
- Notion de mesure textuelle
- Exemple

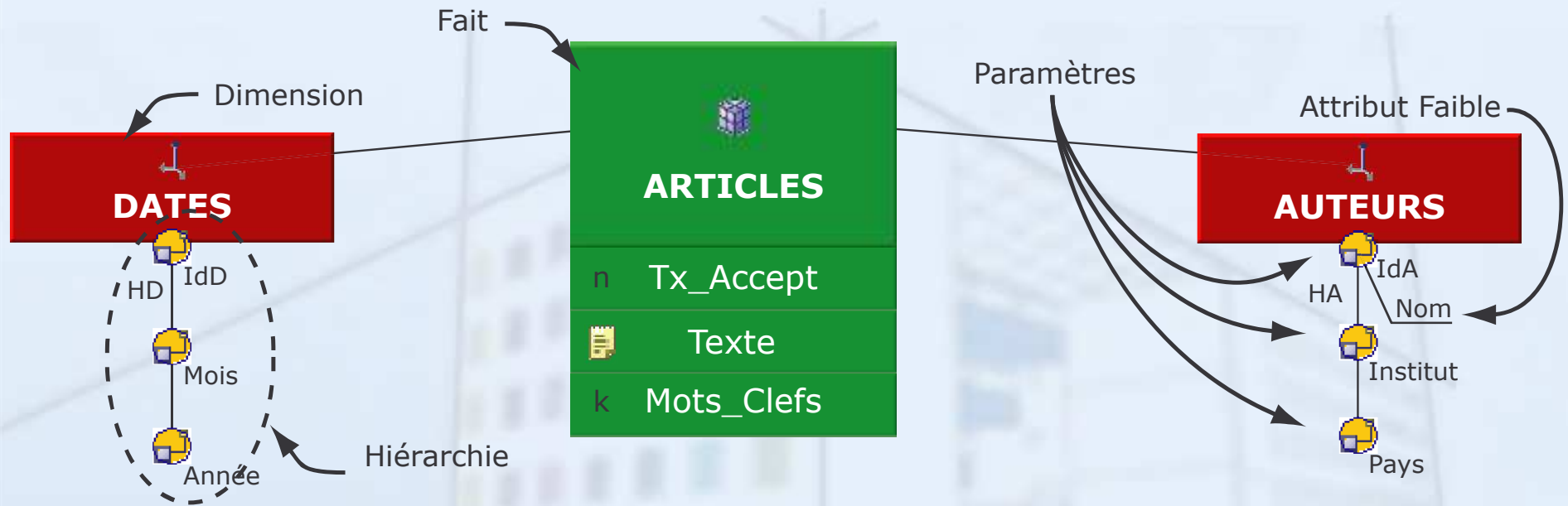
■ Fonction d'agrégation

■ Conclusion





■ Concepts : faits et dimensions



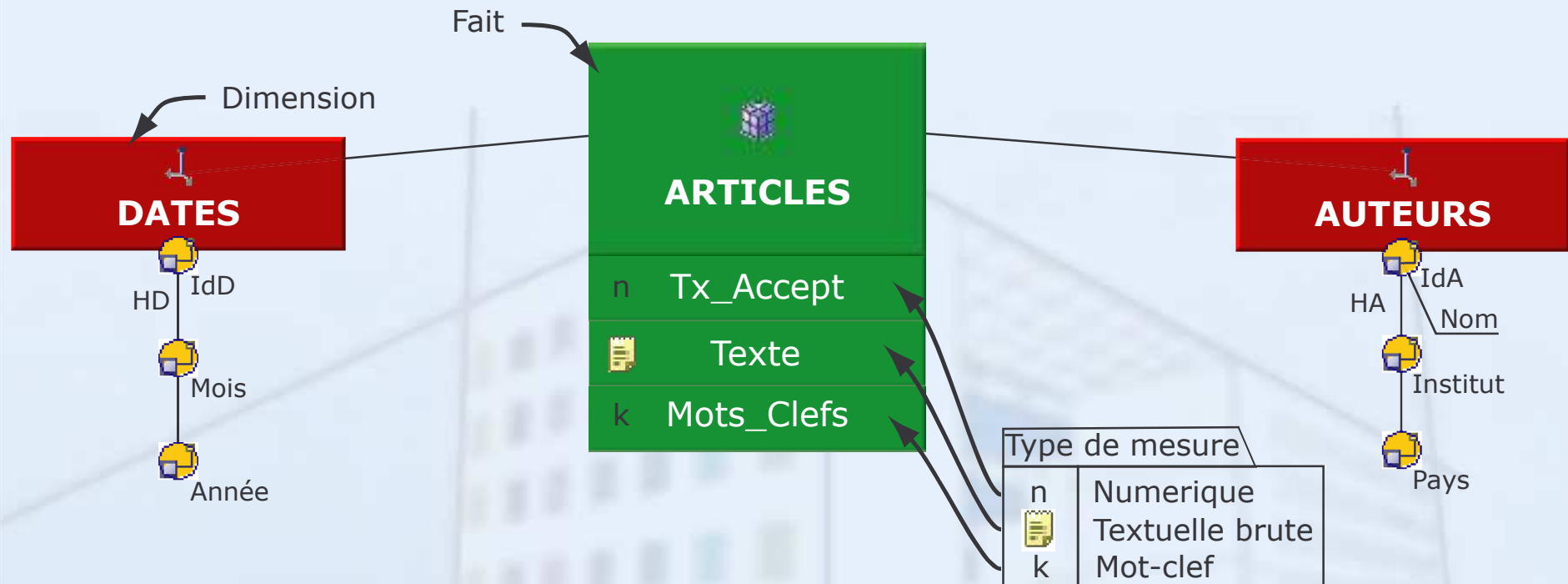


■ Mesures et fonctions d'agrégation

Type de mesure		Fonction applicable	Exemple
Numérique	Additive	Fonction arithmétiques et génériques	quantité d'articles
	Semi-Additive	Avg, Min, Max et génériques	températures
Textuelle	Brute	Top_Kw, génériques	contenu d'articles
	Élaborée (ex. mots-clefs)	Avg_Kw, génériques	mots-clefs d'un fragment de document



Exemple



Analyse de publications en fonction d'auteurs et de la date de publication



- Introduction : contexte
- Modèle conceptuel
- **Fonction d'agrégation**
 - Ordonnancement de termes
 - Fonction d'agrégation
 - Exemple
- Conclusion



■ Calcul d'un poids de « représentativité »

- Emploi d'une fonction de pondération de Recherche d'Information (RI) : *tf.idf*

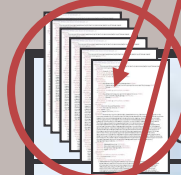
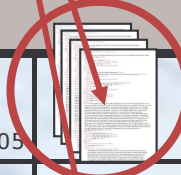
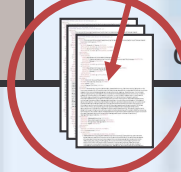
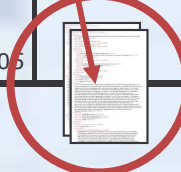
<p>Nb de fois où t est présent dans un fragment</p> <p>Term frequency</p> $tf(t) = \frac{n(t)}{\sum_{frag} n}$ <p>Nb total de termes dans la fragment (normalisation)</p>	<p>Nb total de documents</p> <p>Inverse document frequency</p> $idf(t) = \log \frac{(nb_doc)}{(nb_doc(t))}$ <p>Nb de documents contenant t</p>
---	--

LIMITE : la recherche d'information travaille sur une collection COMPLÈTE



■ Adaptation au décisionnel

- Changement de la notion de collections (RI)

Top_Keyword ₂ (ARTICLES.Texte)		TEMPS		
		Annee	2005	2006
AUTEURS	IdA			
	Au1		 C _{Au1,2005}	 C _{Au1,2006}
	Au2		 C _{Au2,2005}	 C _{Au2,2006}

4 collections à raison d'une par cellule

Analyse des 2 principaux termes d'articles scientifiques
En fonction de l'auteur et de la date de publication

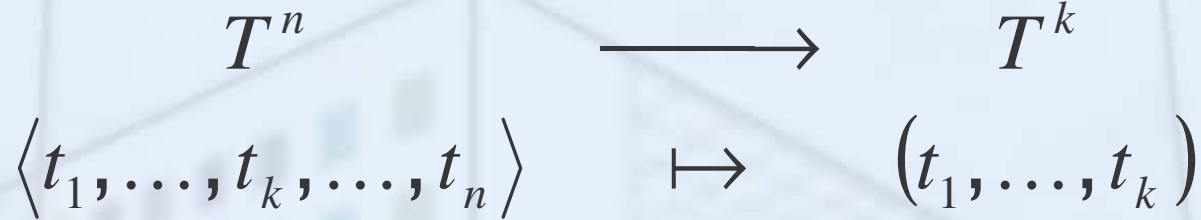




■ Fonction d'agrégation

- Retourne les k termes les plus « représentatifs »

$Top_Kw_k :$



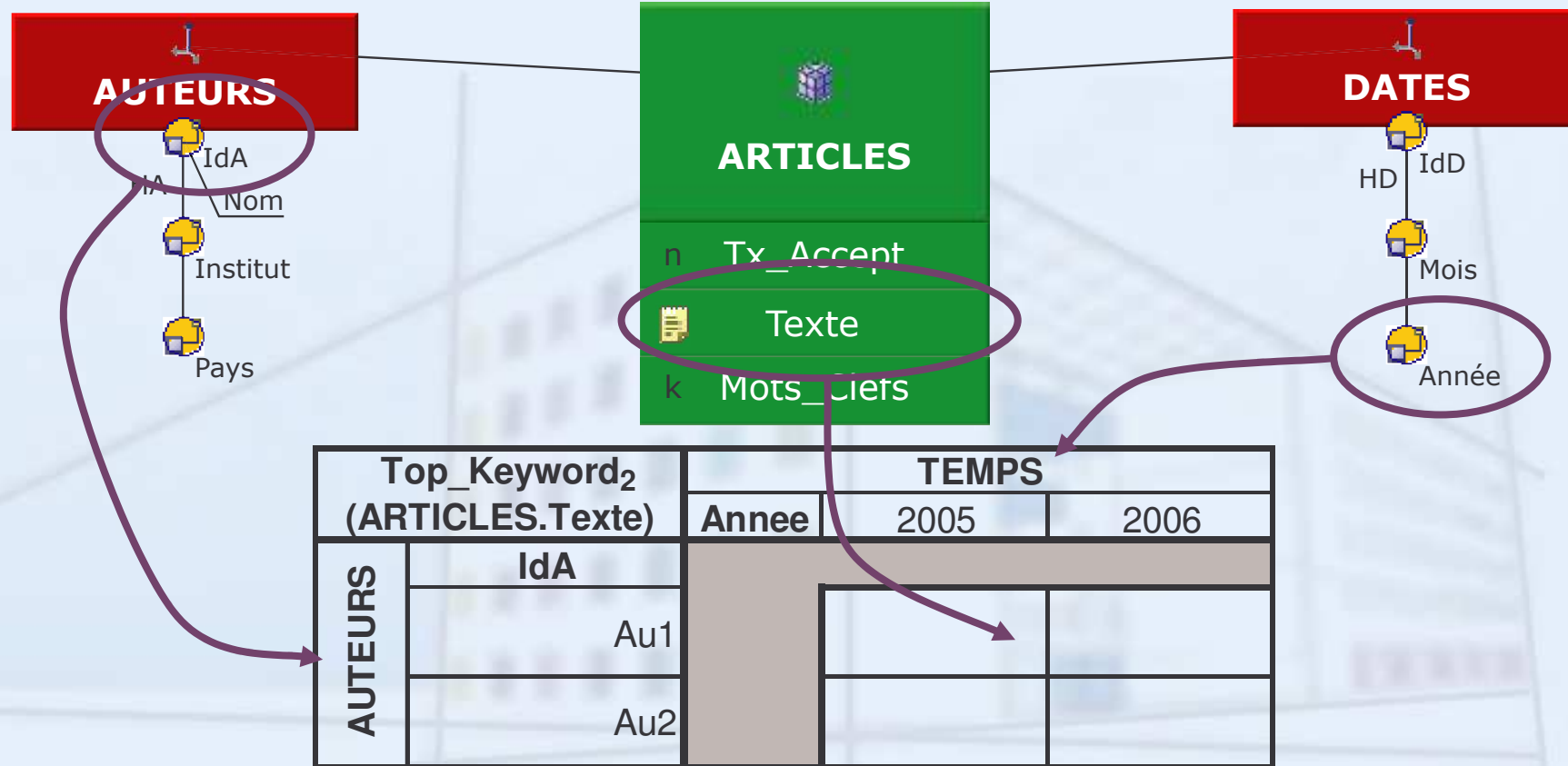
Liste ordonnées
des termes selon
leur poids

Liste des termes
d'une cellule



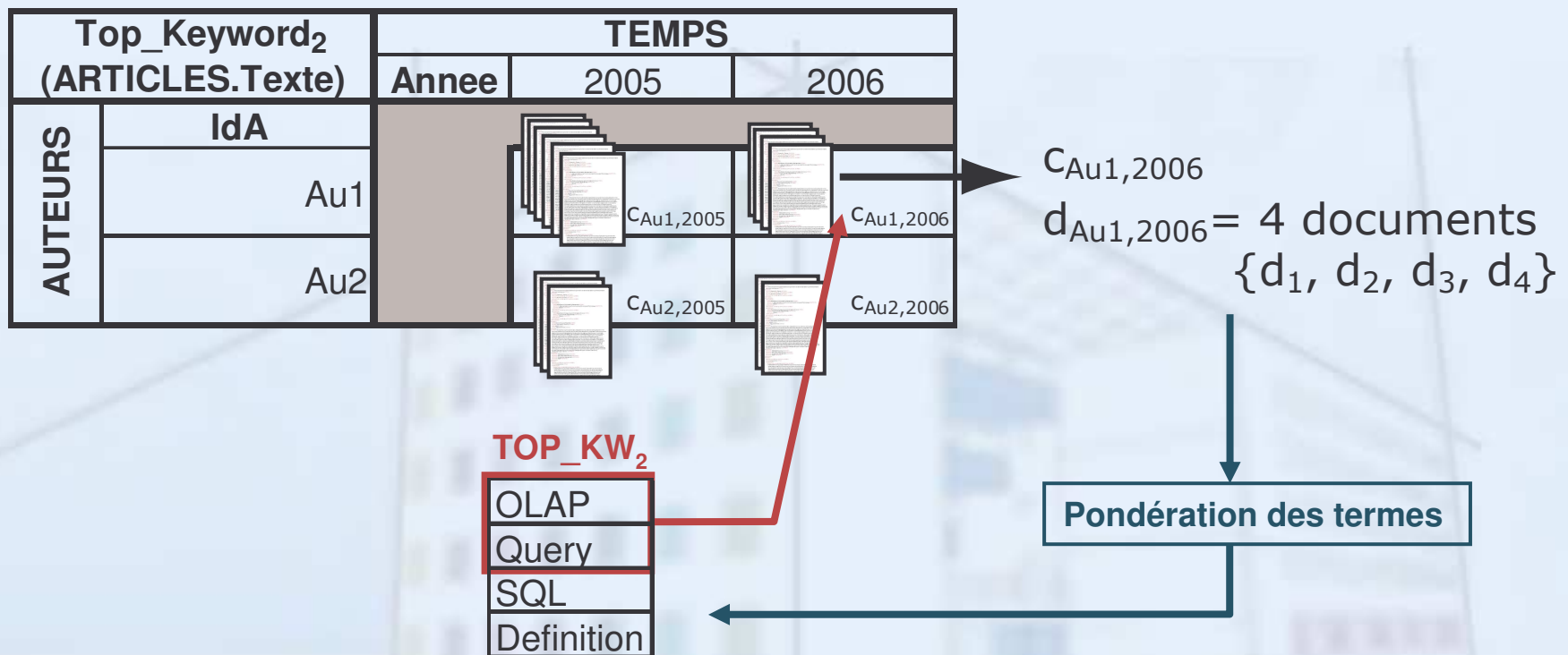
Exemple

Analyse des 2 principaux termes d'articles scientifiques
En fonction de l'auteur et de la date de publication



Exemple

■ Détails de l'agrégation d'une cellule



Exemple



Nb de fois ou le terme est employé

terme	Document			
	d1	d2	d3	d4
OLAP	15	0	18	12
Requete	12	15	0	10
SQL	0	7	10	3
Definition	2	1	16	1

Nombre de docs. avec le terme t

3
3
3
4

idf

0,2218
0,2218
0,2218
0,0969

Nb total de terme dans chaque

2243	1654	1245
------	------	------

OLAP	0,00601817
Requete	0,004304741
SQL	0,003052596
Definition	0,001438742

SOMME

tf

terme	Document			
	d1	d2	d3	d4
OLAP	0,0067	0	0,0145	0,006
Requete	0,0053	0,0091	0	0,005
SQL	0	0,0042	0,008	0,0015
Definition	0,0009	0,0006	0,0129	0,0005

tf.idf

terme	Document			
	d1	d2	d3	d4
OLAP	0.0015	0	0.0032	0.0013
Requete	0,0012	0,002	0	0,0011
SQL	0	0,0009	0,0018	0,0003
Definition	9E-05	6E-05	0,0012	5E-05



Exemple



Top_Keyword ₂ (ARTICLES.Texte)		TEMPS		
		Annee	2005	2006
AUTEURS	IdA			
	Au1		OLAP Entrepôt 2005	OLAP Requetes
	Au2		Entrepôt Document 2005	XML Document 2006





- Introduction : contexte
- Modèle conceptuel
- Fonction d'agrégation
- **Conclusion**
 - Bilan
 - Perspectives



■ Système OLAP avec indicateurs textuels

- Environnement adapté pour des mesures textuelles
- Première approche d'agrégation de mesure textuelle par extraction des termes les plus représentatifs

■ Perspectives

- Alternatives à la fonction de pondération *tf.idf*
- Environnement plus global



■ Merci

EDA'08, Toulouse, Ronan Tournier, tournier@irit.fr

SIG/ED : Systèmes d'Informations Généralisés / Entrepôts de Données

IRIT, Institut de Recherche en Informatique de Toulouse
Université de Toulouse (UT1, UTM, UPS).

