CoStar, un algorithme de co-clustering sans paramètre pour l'analyse de données hétérogènes structurées en étoile[1].

Céline Robardet

LIRIS, UMR 5205, CNRS INSA-Lyon

celine.robardet@insa-lyon.fr

16 février 2012

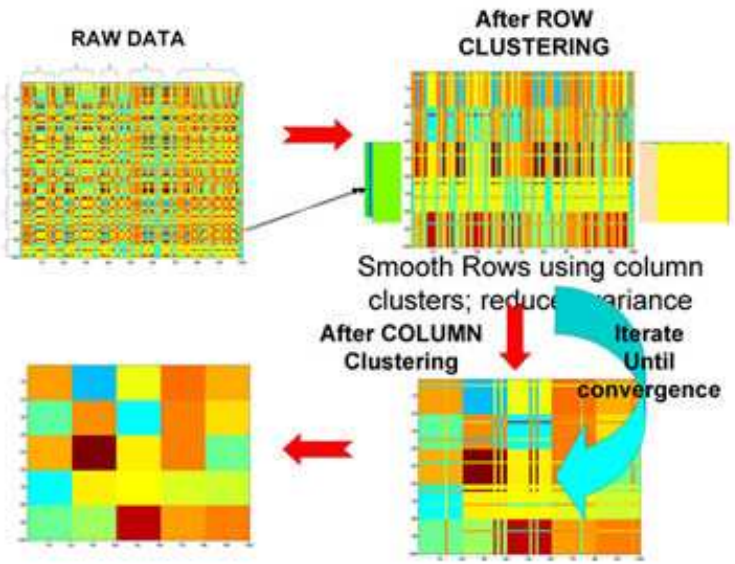
[1] En collaboration avec Dino Ienco, Ruggero G. Pensa et Rosa Meo
Data Mining and Knowledge Discovery (accepté en janvier 2012)

Plan

Motivations

Co-Clustering:

- Partitionne simultanément les lignes et les colonnes d'un tableau de données
- Méthodes efficaces permettant d'obtenir des résultats intéressants
- Utile pour des données de grande dimension



Motivations

Différents types d'algorithmes de Co-clustering :

- Approches spectrales [1]
- Approches basées dur la théorie de l'information [2]
- Approches bayésiennes [3]
- Approches basées sur le χ^2 [4]

[1] I. S. Dhillon, **Co-clustering documents and words using bipartite spectral graph partitioning**, KDD, 2001

[2] I. S. Dhillon et all., **Information-Theoretic Co-Clustering**, KDD, 2003

[3] H. Shan et all., **Bayesian Co-Clustering**, ICDM, 2008

[4] G. Besson et all., **Chi-Sim: A New Similarity Measure for the Co-clustering Task**, ICMLA, 2008

Motivations

Différents types d'algorithmes de Co-clustering :

- Approches spectrales [1]
- Approches basées dur la théorie de l'information [2]
- Approches bayésiennes [3]
- Approches basées sur le χ^2 [4]

Toutes ces méthodes

- Nécessitent le nombre de classes de la partition des lignes et de la partition des colonnes comme paramètres
- Quelques méthodes de co-clustering ont été développées récemment pour l'analyse de données structurées en schéma étoile.

[1] I. S. Dhillon, **Co-clustering documents and words using bipartite spectral graph partitioning**, KDD, 2001

[2] I. S. Dhillon et all., **Information-Theoretic Co-Clustering**, KDD, 2003

[3] H. Shan et all., **Bayesian Co-Clustering**, ICDM, 2008

[4] G. Besson et all., **Chi-Sim: A New Similarity Measure for the Co-clustering Task**, ICMLA, 2008

Considérations générales - 1

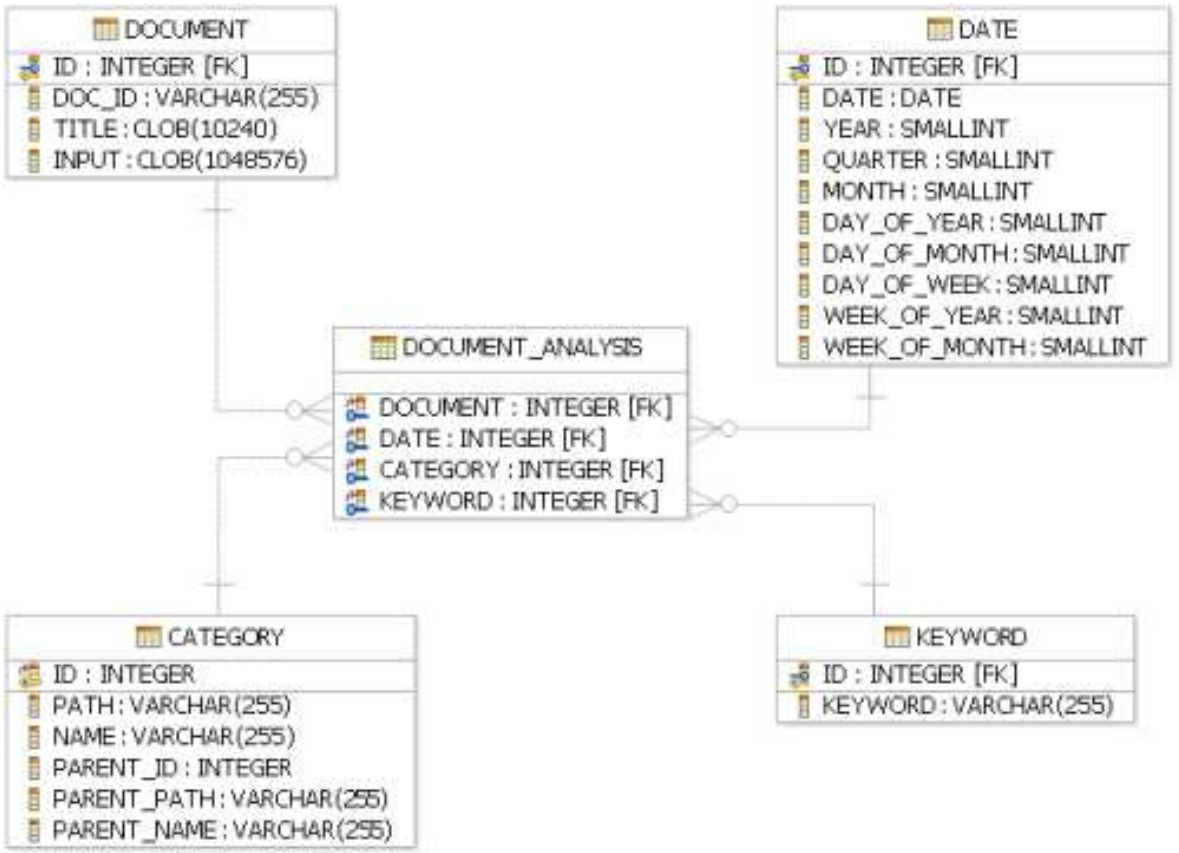
Le clustering est souvent utilisé pour obtenir une première vue schématique des données
Habituellement les données contiennent une seule vue

	Children's use			Mothers' use		
Emotion word	Girls	Boys	Total	Girls	Boys	Total
Afraid	0	2	2	2	4	6
Enjoy	0	0	0	1	6	7
Excited	1	0	1	11	4	15
Feel better	4	0	4	5	7	12
Feel good	2	0	2	2	1	3
Feel hurt	0	0	0	4	1	5
Fun[a]	3	5	8	71	36	107
Glad[a]	2	0	2	32	26	58
Happy[a]	7	6	13	34	77	111
Like[a]	15	16	31	51	48	99
Love[a]	13	2	15	23	26	49
Mad	3	3	6	13	11	24
Missed[a]	7	2	9	73	57	130
Sad[a]	20	10	30	38	71	109
Scared	3	0	3	9	15	24
Sorry[a]	5	8	13	8	10	18
Surprised	2	5	7	6	6	12
Upset	0	0	0	8	7	15
Worried	1	6	7	19	9	28

$\Rightarrow$ Le nombre de classes est un paramètre difficile à déterminer

Considérations générales - 2

Mais les données structurées en schéma étoile permettent d'avoir plus d'information



Elles contiennent plusieurs vues, définies par des ensembles de descripteurs différents des mêmes objets.

Notre idée

Approche proposée

- Co-Clustering : permet de mettre en évidence les relations entre objets et descripteurs
- Construction simultanée d'un co-clustering sur chaque vue :
  - chaque vue permet la construction d'une partie du modèle
  - les différentes représentations des objets mais aussi les modèles intermédiaires sur chaque vue sont utilisés pour construire les modèles des autres vues.

Adaptation de la méthode de co-clustering τ-CoClust [1] aux données structurées en étoile

$\Rightarrow$ Permet d'obtenir une méthode de co-clustering pour données structurées en schéma étoile sans aucun paramètre.

[1] C. Robardet, PhDThesis, **Contribution à la classification non supervisée: proposition d'une methode de bi-partitionnement**, Université Claude Bernard - Lyon 1, 2002

τ-CoClust méthode

Quelques caractéristiques

- Co-Clustering méthode pour des données de type catégorielle (présence/absence) ou fréquencielle
- Pas de paramètre à fixer par l'utilisateur
- Maximise le mesure statistique τ de Goodman et Kruskal entre les partitions des lignes et des colonnes
- Produit un indicateur de qualité de chaque partition.

La mesure τ de Goodman et Kruskal

Étant donné un tableau de données D et un co-clustering $\{CO, CF\}$, on calcule un tableau de contingence T

		CF_1		CF_2	
		f_1	f_2	f_3	f_4
CO_1	o_1	3	4	1	1
	o_2	5	3	0	2
	o_3	6	4	1	0
CO_2	o_4	0	1	7	7
	o_5	1	0	6	8

Tableau de données D

$$T_{ij} = \sum_{o_k \in CO_i} \sum_{f_\ell \in CF_j} D[k, \ell]$$

Tableau de contingence T

La mesure $\tau_{CO|CF}$ de Goodman et Kruskal évalue la réduction proportionnelle de l'erreur dans la prédiction de la variable dépendante CO due à la prise en compte de la variable indépendante CF :

- e_{CO} : Erreur de prédiction sur CO lorsque l'on ne connaît pas CF
- $E[e_{CO|CF}]$: Espérance de l'erreur de prédiction sur CO lorsque l'on connaît CF.

$$\tau_{CO|CF} = \frac{e_{CO} - E[e_{CO|CF}]}{e_{CO}}$$

La mesure τ de Goodman et Kruskal : quelques propriétés

τ satisfait plusieurs propriétés souhaitables d'une mesure de co-clustering :

- Elle est invariante par permutation des lignes et des colonnes.
- Elle prend ses valeurs dans $[0, 1]$
- Elle a une signification opérationnelle
- Contrairement à d'autres mesures d'association (χ^2) τ a une limite supérieure qui est définie indépendamment des nombres de classes. Ainsi, τ peut être utilisée pour comparer des co-clusterings de différentes tailles.

Mesure τ de Goodman et Kruskal : un exemple

	f_1	f_2	f_3	f_4
o_1	3	4	1	1
o_2	5	3	0	2
o_3	6	4	1	0
o_4	0	1	7	7
o_5	1	0	6	8

$CO = \{\{o_1, o_2, o_3\}, \{o_4, o_5\}\}$
$CF = \{\{f_1, f_2\}, \{f_3, f_4\}\}$

$$T_{ij} = \sum_{o_k \in CO_i} \sum_{f_\ell \in CF_j} D[k, \ell]$$

Mesure τ de Goodman et Kruskal : un exemple

	f_1	f_2	f_3	f_4
o_1	3	4	1	1
o_2	5	3	0	2
o_3	6	4	1	0
o_4	0	1	7	7
o_5	1	0	6	8

$CO = \{\{o_1, o_2, o_3\}, \{o_4, o_5\}\}$
$CF = \{\{f_1, f_2\}, \{f_3, f_4\}\}$

$$T_{ij} = \sum_{o_k \in CO_i} \sum_{f_\ell \in CF_j} D[k, \ell]$$

	CF_1	CF_2
CO_1	25	5
CO_2	2	28

Mesure τ de Goodman et Kruskal : un exemple

(A)

	f_1	f_2	f_3	f_4
o_1	3	4	1	1
o_2	5	3	0	2
o_3	6	4	1	0
o_4	0	1	7	7
o_5	1	0	6	8

	CF_1	CF_2	
CO_1	25	5	30
CO_2	2	28	30
	27	33	60

(B)

	f_1	f_2	f_3	f_4
o_1	3	4	1	1
o_2	5	3	0	2
o_3	6	4	1	0
o_4	0	1	7	7
o_5	1	0	6	8

	CF_1	CF_2	
CO_1	15	4	19
CO_2	12	29	41
	27	33	60

Nous obtenons $\tau_{CO|CF} = 0.5937$ pour le co-clustering (A) et $\tau_{CO|CF} = 0.21577$ pour le co-clustering (B).

Stratégie d'optimisation

- La mesure τ n'est pas symétrique : l'algorithme optimise alternativement chacune des deux fonctions $\tau_{CO|CF}$ et $\tau_{CF|CO}$

Stratégie d'optimisation

- La mesure τ n'est pas symétrique : l'algorithme optimise alternativement chacune des deux fonctions $\tau_{CO|CF}$ et $\tau_{CF|CO}$

Algorithme τ-CoClust : Optimisation stochastique de CO et CF

Initialisation de CO et CF
Pour $i = 1$ to nb_iter **Faire**
  #Optmisation de CO
  Prendre aléatoirement $CO_b \in CO$
  Prendre aléatoirement $o \in CO_b$
  $max_{\tau_{CO|CF}} \leftarrow \tau_{CO|CF}$
  $CO' \leftarrow CO$
  Pour tout $CO_e \in \{CO \cup \emptyset\}$ **Faire**
    $CO'_b \leftarrow CO_b \setminus \{o\}$
    $CO'_e \leftarrow CO_e \cup \{o\}$
    Si$(\tau_{CO'|CF} > max_{\tau_{CO|CF}})$ **alors**
      $max_{\tau_{CO|CF}} \leftarrow \tau_{CO'|CF}$
    Fin Si
  Fin Pour
  #Optmisation de CF
  Prendre aléatoirement $CF_b \in CF$
  Prendre aléatoirement $f \in CF_b$
  $max_{\tau_{CF|CO}} \leftarrow \tau_{CF|CO}$
  $CF' \leftarrow CF$
  Pour tout $CF_e \in \{CF \cup \emptyset\}$ **Faire**
    $CF'_b \leftarrow CF_b \setminus \{f\}$
    $CF'_e \leftarrow CF_e \cup \{f\}$
    Si$(\tau_{CF'|CO} > max_{\tau_{CF|CO}})$ **alors**
      $max_{\tau_{CF|CO}} \leftarrow \tau_{CF'|CO}$
    Fin Si
  Fin Pour
Fin Pour

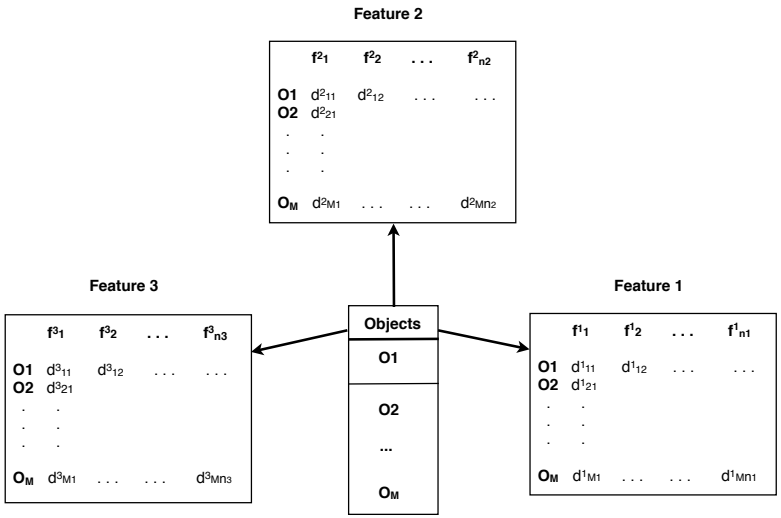
CoStar méthode

- Généralisation de τ-CoClust aux données structurées en schéma étoile
- Extension de la mesure τ de Goodman et Kruskal aux données étoile

$\Rightarrow$ On a $N+1$ fonctions objectives à maximiser, où N est le nombre de vues

- Adaptation de la stratégie d'optimisation locale afin de garantir l'obtention d'un maximum de Pareto sur cet ensemble de fonctions.

Mesure τ pour les données en schéma étoile



Mesures τ de Goodman et Kruskal généralisées aux données en schéma étoile :

$$\tau_{CO|\{CF^1,\ldots,CF^N\}} = \frac{\sum_{n=1}^{N} e_{CO^n} - \sum_{n=1}^{N} E[e_{CO^n|CF^n}]}{\sum_{n=1}^{N} e_{CO^n}}$$

et

$$\tau_{CF^n|CO} = \frac{e_{CF^n} - E[e_{CF^n|CO^n}]}{e_{CF^n}}$$

Mesure τ pour les données en schéma étoile : un exemple

	f_1^1	f_2^1	f_3^1	f_4^1	f_1^2	f_2^2	f_3^2
o_1	3	4	1	1	0	8	5
o_2	5	3	0	2	0	6	9
o_3	6	4	1	0	2	2	2
o_4	0	1	7	7	9	1	0
o_5	1	0	6	8	7	0	1

	CF_1^1	CF_2^1		CF_1^2	CF_2^2	
CO_1	25	5	30	2	32	34
CO_2	2	28	30	16	2	18
	27	33	60	18	34	52

$$\tau_{CO|\{CF^1,CF^2\}} = 0.6390,\ \tau_{CF^1|CO} = 0.5937,\ \tau_{CF^2|CO} = 0.6890$$

Mesure τ pour les données en schéma étoile : un exemple

	f_1^1	f_2^1	f_3^1	f_4^1	f_1^2	f_2^2	f_3^2
o_1	3	4	1	1	0	8	5
o_2	5	3	0	2	0	6	9
o_3	6	4	1	0	2	2	2
o_4	0	1	7	7	9	1	0
o_5	1	0	6	8	7	0	1

	CF_1^1	CF_2^1		CF_1^2	CF_2^2	
CO_1	25	5	30	2	32	34
CO_2	2	28	30	16	2	18
	27	33	60	18	34	52

$$\tau_{CO|\{CF^1,CF^2\}} = 0.6390,\ \tau_{CF^1|CO} = 0.5937,\ \tau_{CF^2|CO} = 0.6890$$

	f_1^1	f_2^1	f_3^1	f_4^1	f_1^2	f_2^2	f_3^2
o_1	3	4	1	1	0	8	5
o_2	5	3	0	2	0	6	9
o_3	6	4	1	0	2	2	2
o_4	0	1	7	7	9	1	0
o_5	1	0	6	8	7	0	1

	CF_1^1	CF_2^1		CF_1^2	CF_2^2	
CO_1	18	17	35	9	18	27
CO_2	12	13	25	9	16	25
	30	30	60	18	34	52

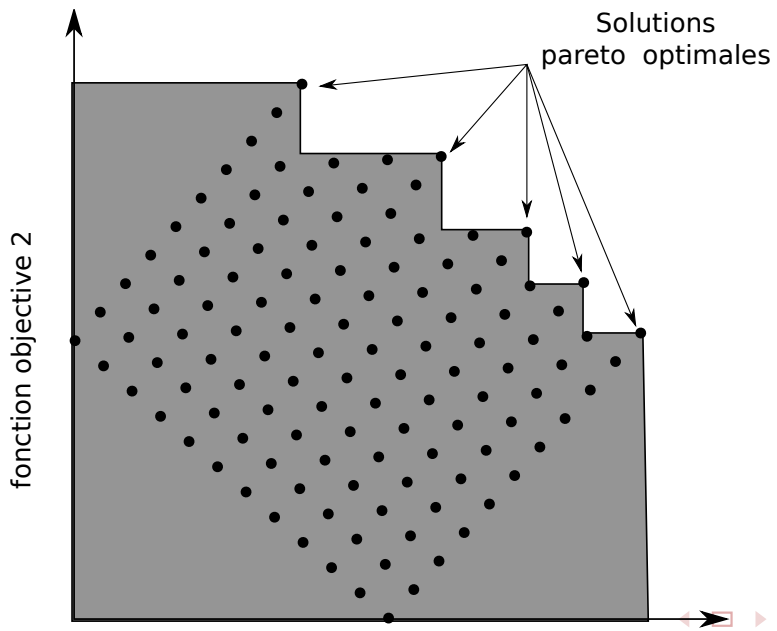
$$\tau_{CO|\{CF^1,CF^2\}} = 0.0119,\ \tau_{CF^1|CO} = 0.0234,\ \tau_{CF^2|CO} = 0.0008$$

CoStar : stratégie d'optimisation des $N+1$ mesures τ

Optimisation multi-objectif :

$$\max_{CC=(CO,CF^1,\ldots,CF^N)} \tau(CC) = \left(\tau_{CF^1|CO}, \ldots, \tau_{CF^N|CO}, \tau_{CO|\{CF^1,\ldots,CF^N\}}\right)$$

Une solution est Pareto-optimale si aucune amélioration ne peut être faite sur une fonction objectif sans dégradation sur au moins une autre fonction.



CoStar : l'algorithme

Algorithme CoStar : Optimisation stochastique de CO et $\{CF^1, \ldots, CF^N\}$

Initialisation de CO et $CF^1, \ldots, CF^N$
Pour $i = 1$ to nb_iter **Faire**
    Optmisation(CO, $\tau_{CO|\{CF^1,\ldots,CF^N\}}$)
    Pour $n = 1$ à N **Faire** Optmisation(CF^n, $\tau_{CF^n|CO}$)
Fin Pour

CoStar : l'algorithme

Algorithme CoStar : Optimisation stochastique de CO et $\{CF^1, \ldots, CF^N\}$

Initialisation de CO et $CF^1, \ldots, CF^N$
Pour $i = 1$ to nb_iter **Faire**
  Optmisation(CO, $\tau_{CO|\{CF^1,\ldots,CF^N\}}$)
  Pour $n = 1$ à N **Faire** Optmisation(CF^n, $\tau_{CF^n|CO}$)
Fin Pour

Algorithme Optimisation (C, τ_C) : retourne $Pareto$, une partition Pareto-optimale dans le voisinage de C.

$max \leftarrow \tau_C(C)$
Pour tout C' dans le voisinage de C **Faire**
  Si $\tau_C(C') > max$ **Alors**
    $max \leftarrow \tau_C(C')$; $Pareto \leftarrow \{C'\}$
  Fin Si
  Si $\tau_C(C') = max$ **Alors**
    $Pareto \leftarrow Pareto \cup \{C'\}$
  Fin Si
Fin Pour

Tant que $\#Pareto > 1$ **Faire**
  Prendre C_1 et C_2 dans Pareto
  $n \leftarrow 1$
  Tant que $(\tau(CC_1)_n = \tau(CC_2)_n)$
    et $(n < N)$ **Faire** $n \leftarrow n + 1$
  Si $\tau(CC_1)_n < \tau(CC_2)_n$ **Alors**
    $Pareto \leftarrow Pareto \setminus \{C_1\}$
  Sinon
    $Pareto \leftarrow Pareto \setminus \{C_2\}$
  Fin Si
Fin Tant que

CoStar : caractéristiques

- CoStar converge localement vers un co-clustering Pareto-optimal.
- La complexité de CoStar est réduite grâce à un calcul incrémental de la valeur des fonctions objectifs :

$$O(nb\_iter \times \#Objets \times (\sum_{n=1}^{N} \#F^n))$$

Expérimentation : les compétiteurs

- Comparaison avec 3 algorithmes de co-clustering pour données structurées en étoile :
  - COMRAF[1] : Combinatorial Markov Random Field
  - SRC[2] : Spectral Relational Clustering
  - NMF[3] : Semi-Supervised Non-negative Matrix Factorization

[1] Bekkerman, R., Jeon, J., **Multi-modal clustering for multimedia collections**, Computer Vision and Pattern Recognition, 2007

[2] B. Long et al.**Spectral clustering for multi-type relational data**, ICML, 2006

[3] Y. Chen et al. **Semi-supervised document clustering with simultaneous text representation and categorization**, ECML/PKDD, 2009

Expérimentation : les données

- Des jeux de données synthétiques public (BBC et BBCSport)
- Trois bases de données document-terms-categories pour illustrer les cas où les dimensions sont déséquilibrées
- Un jeu de données image-terms-catégorie (COREL Benchmark) pour illustrer les problèmes équilibrés

Expérimentation : données BCC et BCCSport

- Jeux de données construits à partir d'articles de presse.
- Les classes d'objets correspondent aux annotations des articles de presse (affaires, divertissement, politique, sport, technologie).

Name	# obj.	# obj. classes	View	# features
bbc2	2012	5	1st view	6838
			2nd view	6790
bbcsport2	544	5	1st view	3183
			2st view	3203
bbc4	685	5	1st view	4659
			2nd view	4633
			3rd view	4665
			4th view	4684

Expérimentation : données Document-Terms-Categories

Name	# terms	# cat.	# doc. (obj.)	# classes	Classes
T1	3987	2	833	5	{Aden-Diph, Cell-Mov, Aluminium} {cpi, money }
T2	3197	2	461	5	{Blood-Coag, Enzyme-Act, Staph-Inf} {jobs, reserves}
T3	8282	3	2129	9	{ Film, Television, Health} {Aden-Diph, Cell-Mov, Enzyme-Act} { interest, trade, money}

Caractéristiques des jeux de données textuelles.

Class Name	# obj.	Class Name	# obj.
Aden-Diph	56	*Cpi*	60
Cell-Mov	106	*Money*	608
Aluminium	53	*Interest*	219
Blood-Coag	69	*Trade*	319
Enzyme-Act	154	*Film*	196
Staph-Inf	157	*Television*	130
Jobs	39	*Health*	341
Reserves	42		

Distribution des classes pour les jeux de données textuelles.

Expérimentation : données Image-Terms-Categorie

Name	# terms	# images (obj.)	# catégories	Classes
I1	114	630	7	sunsets, tigers, train, swimmers, formula One car, skyscrapers, war airplanes
I2	116	541	6	bears, deers, horses, cliffs, birds, bridges
I3	170	1171	13	sunsets, tigers, train, swimmers, formula One car, skyscrapers, war airplanes, bridges, bears, deers, horses, cliffs, birds,

Caractéristiques des jeux de données image.

Expérimentation : les mesures d'évaluation

Deux indices de validation externe qui évalue comment la partition calculée est en adéquation avec la variable de classe des données :

- Information Mutuelle Normalisée (NMI)
- Index de Rand Ajusté (ARI)

Expérimentation : comparaison avec les compétiteurs

	CoStar	Comraf	SRC	NMF
bbc2	**0.68** ± 0.02	0.58 ± 0.04	0.31 ± 0.11	0.5 ± 0.04
bbcsport2	**0.69** ± 0.06	0.09 ± 0.05	0.24 ± 0	0.57 ± 0.01
bbc4	**0.68** ± 0.03	0.41 ± 0.03	0.33 ± 0.04	0.45 ± 0.0
l1	**0.88** ± 0.04	0.85 ± 0.05	0.80 ± 0.11	0.86 ± 0.06
l2	**0.75** ± 0.04	0.66 ± 0.05	0.61 ± 0.09	0.74 ± 0.04
l3	**0.76** ± 0.03	0.75 ± 0.04	0.61 ± 0.11	0.73 ± 0.06
T1	**0.72** ± 0	0.50 ± 0.03	0.23 ± 0.05	0.49 ± 0.01
T2	**0.73** ± 0.1	0.35 ± 0.03	0.33 ± 0.08	0.58 ± 0.02
T3	**0.71** ± 0.02	0.66 ± 0	0.55 ± 0.06	0.63 ± 0.02

Comparaison selon la mesure NMI.

Expérimentation : comparaison avec les compétiteurs

	CoStar	Comraf	SRC	NMF
bbc2	**0.67** ± 0.03	0.42 ± 0.05	0.18 ± 0.08	0.48 ± 0.06
bbcsport2	**0.58** ± 0.14	0.23 ± 0.04	0.14 ± 0.09	0.56 ± 0.02
bbc4	**0.56** ± 0.2	0.34 ± 0.05	0.16 ± 0.06	0.41 ± 0.0
l1	**0.83** ± 0.09	0.80 ± 0.08	0.69 ± 0.11	0.79 ± 0.1
l2	**0.72** ± 0.07	0.55 ± 0.06	0.51 ± 0.09	0.67 ± 0.07
l3	0.50 ± 0.11	**0.62** ± 0.06	0.43 ± 0.13	0.61 ± 0.09
T1	**0.73** ± 0	0.23 ± 0.05	0.09 ± 0.09	0.23 ± 0.02
T2	**0.68** ± 0	0.19 ± 0.04	0.18± 0.09	0.48 ± 0.01
T3	**0.44** ± 0.02	0.42 ± 0.03	0.36 ± 0.08	0.44 ± 0.04

Comparaison selon la mesure ARI.

Expérimentation : validation statistique des résultats

- Test de Friedman : les classement obtenus par les méthodes sont significativement différents
- Test de Nemenyi : la performance de deux méthodes est significativement différente si les rangs moyens de ces méthodes diffèrent d'au moins 1.394.

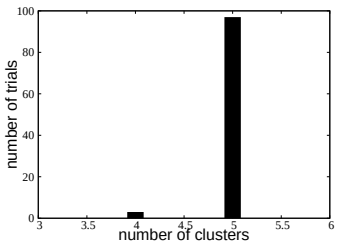
	Average ranking	
Methods	**ARI**	**NMI**
CoStar	1.2	1
NMF	2.2	2.4
ComRaf	2.6	2.7
SRC	4	3.9

ARI	*NMF*	*ComRaf*	*SRC*
CoStar	1	**1.4**	**2.8**
NMF	-	0.4	**1.8**
ComRaf	-	-	**1.4**
SRC	-	-	0

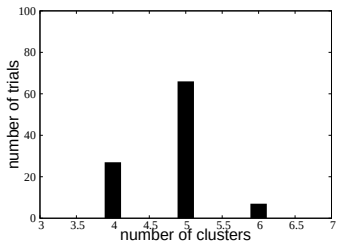
NMI	*NMF*	*ComRaf*	*SRC*
CoStar	**1.4**	**1.7**	**2.9**
NMF	-	0.3	**1.5**
ComRaf	-	-	1.2
SRC	-	-	0

Moyenne des rangs des résultats pour le test de Nemenyi (haut), Différence entre les moyennes des rangs pour les mesures ARI et NMI (bas).

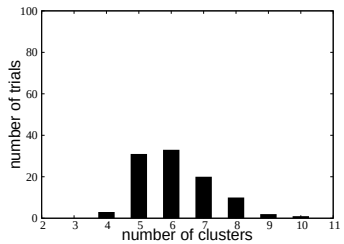
Expérimentation : étude de la variabilité du nombre de classes



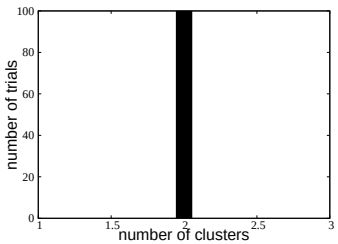
bbc2



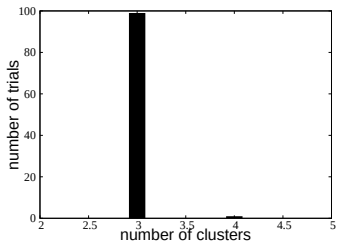
bbcsport2



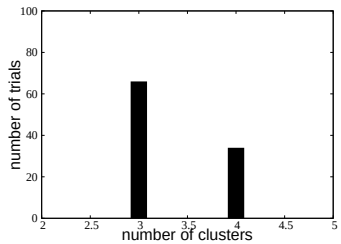
bbc4



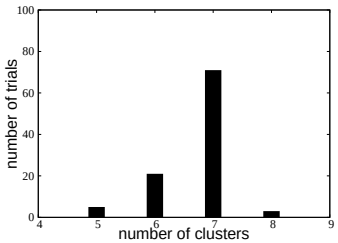
T1



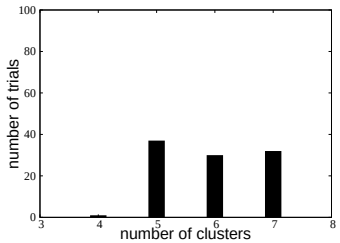
T2



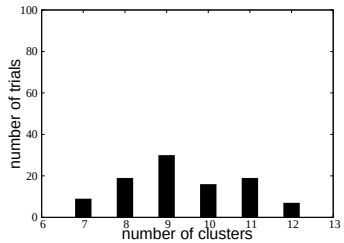
T3



I1

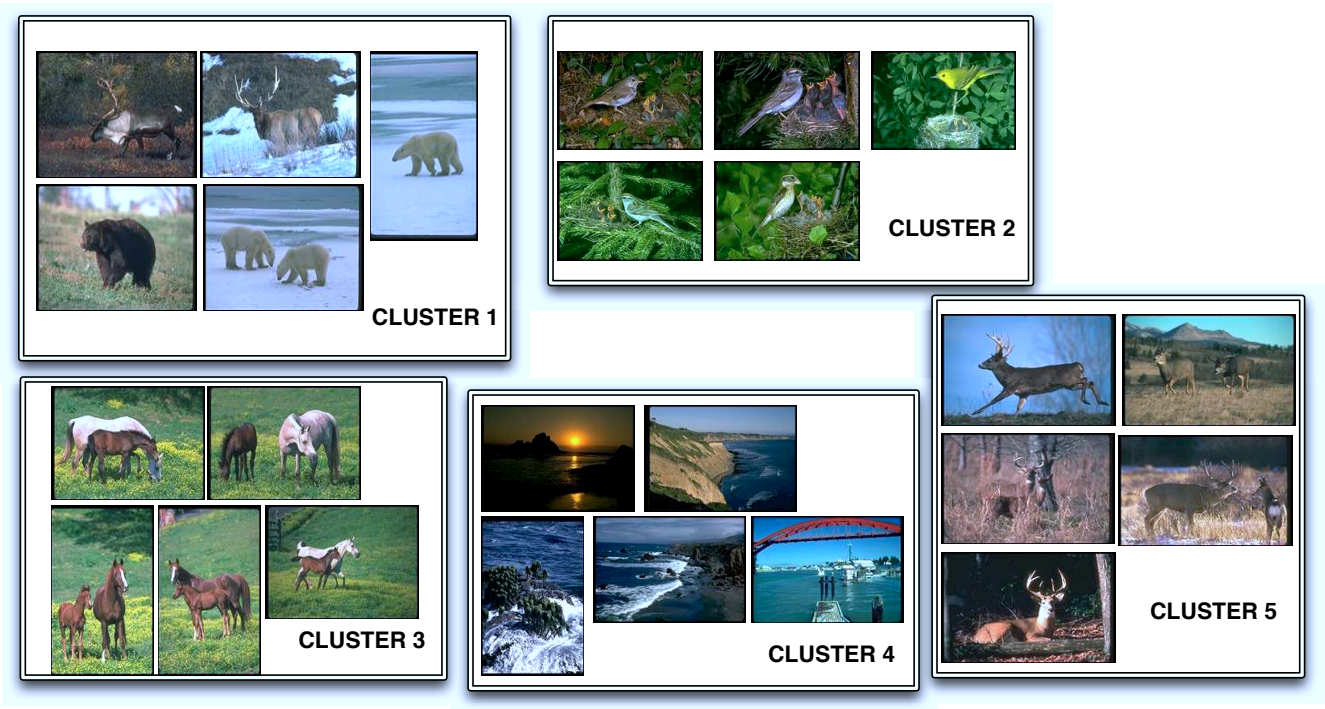


I2



I3

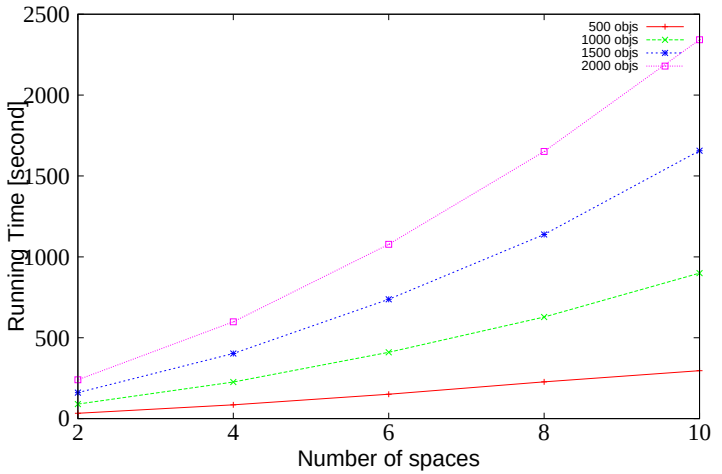
Expérimentation : analyse qualitative

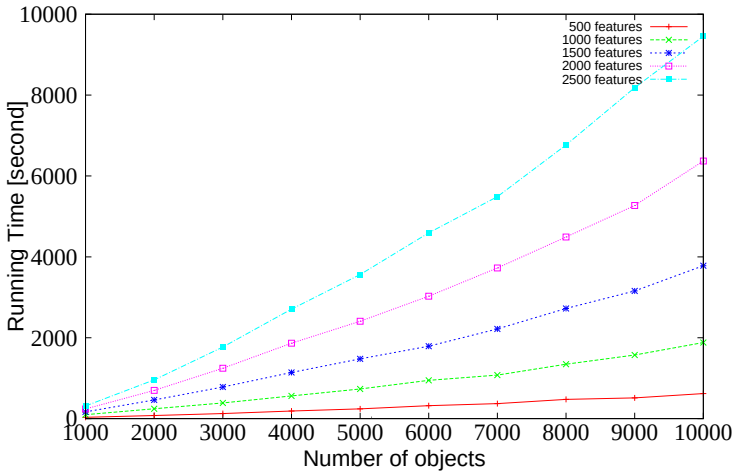


cluster 1	bear, polar, black, snow, antlers, grizzly, elk, ice, tundra, grass
cluster 2	bridge, water, coast, arch, hills, sky, waves, beach, boats, steel
cluster 3	deer, white-tailed, mule, horns, slope, fawn
cluster 4	horses, foals, mare, field, fence, bush
cluster 5	birds, nest, branch, fly, tree, leaf, wood, wings, stick, baby

Données I_2 : 5 images les plus représentatives de chaque classe et terms associés.

Expérimentation : temps d'exécution





Temps d'exécution de CoStar lorsque le nombre de vues et le nombre d'objets varient.

Conclusion

Nous avons proposé CoStar, un algorithme qui

1. produit un co-clustering pour des données structurées en schéma étoile
2. ne nécessite aucun paramètre
3. converge localement vers une solution Pareto-optimale

Conclusion

Nous avons proposé CoStar, un algorithme qui

1. produit un co-clustering pour des données structurées en schéma étoile
2. ne nécessite aucun paramètre
3. converge localement vers une solution Pareto-optimale

Nous avons empiriquement montré que CoStar

1. permet d'obtenir de bons résultats quantitatifs, selon des mesures objectives (indexes de validation externe de partitions)
2. produit de meilleurs résultats, validés statistiquement, que 3 compétiteurs
3. produit des résultats qualitatifs qui semblent pertinents

Merci! :-)

... des questions ?