Analyse des données évolutives :
application aux données d'usage du Web

Alzennyr GOMES DA SILVA

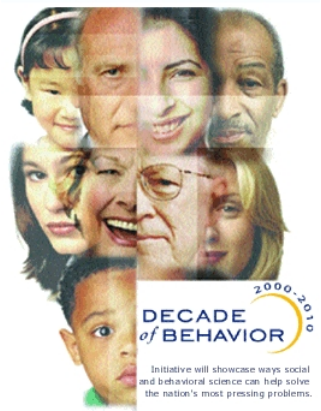
Directeur de thèse : Pr Edwin DIDAY
Co-directeur de thèse : Dr Yves LECHEVALLIER

24 septembre 2009

Plan

Contexte/Motivation

- Le Web : source de données volumineuses et dynamiques



- L'accès aux pages Web a une nature dynamique
  - Changement du contenu et de la structure du site Web
  - Changement d'intérêt/comportement des utilisateurs

- Le comportement d'accès d'un internaute peut dépendre :
  - de l'heure et du jour de la semaine
  - des événements saisonniers
  - des événements ponctuels (crises, catastrophes, compétitions sportives)
  - etc.

Problématique générale

La prise en compte de la dimension temporelle dans l'analyse des données évolutives.

Travaux existants dans le cadre de l'analyse des données évolutives

- **Approches d'extraction de motifs séquentiels**
  (Masseglia *et al.*, 2004; Baron & Spiliopoulou, 2001; Chen & Petrounias, 1999; Liu *et al.*, 2001), etc.

- **Systèmes d'aide à la navigation basés sur le RàPC**
  BRODWAY (Jaczynski & Trousse, 1999), RADIX (Corvaisier *et al.*, 1997), CASEP2 (Zehraoui *et al.*, 2004), COBRA (Malek & Kanawati, 2001), etc.

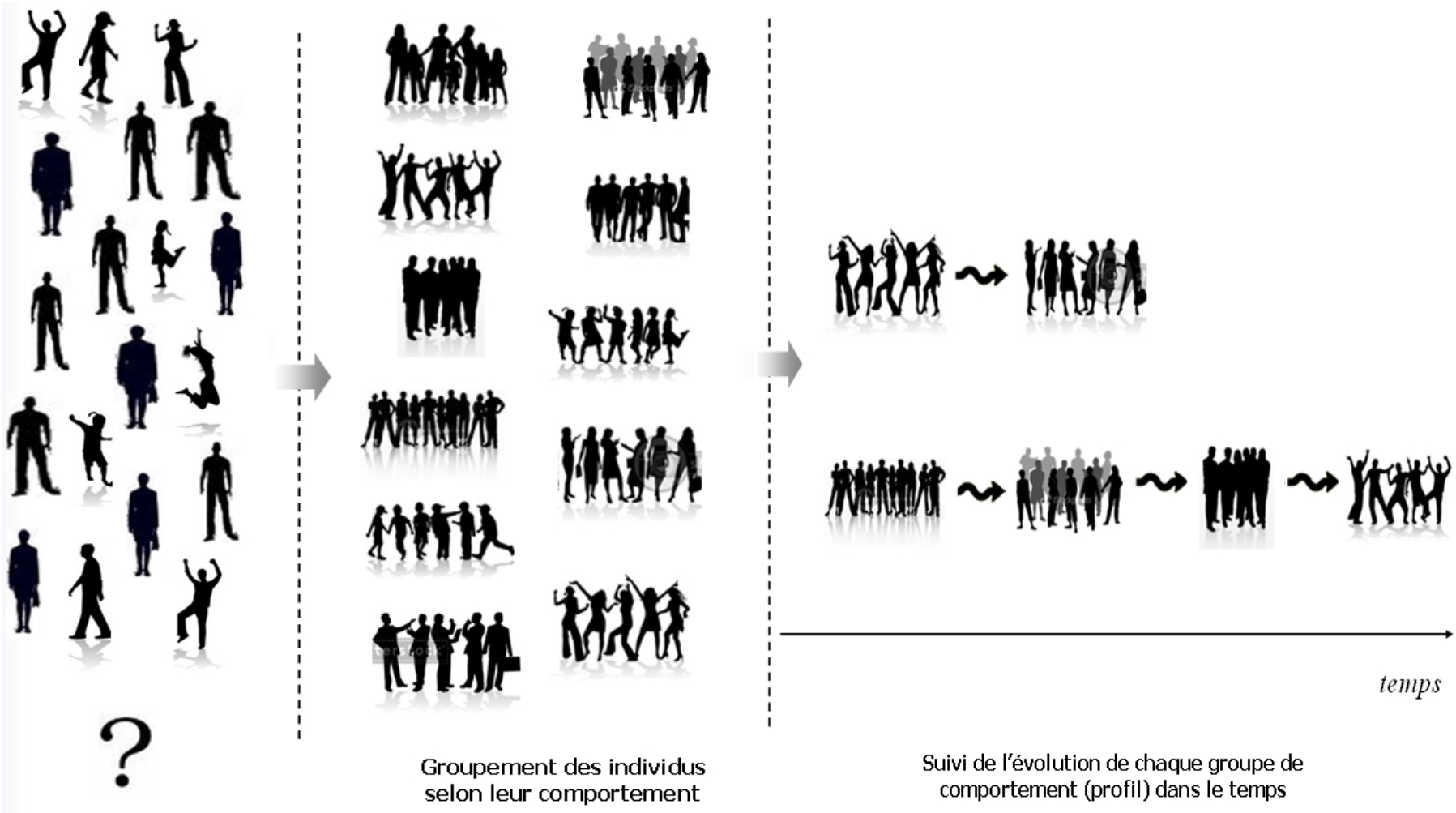
- **Systèmes de détection de changements basés sur le clustering**
  FOCUS (Ganti *et al.*, 1999), DEMON (Ganti *et al.*, 2000), PANDA (Bartolini *et al.*, 2004), MONIC (Spiliopoulou *et al.*, 2006), etc.

- **Approches basées sur le clustering spatio-temporel**
  (Neill *et al.*, 2005), (Aggarwal, 2005), etc.

- **Approches basées sur l'Analyse des Données Symboliques (ADS)**
  Temporal Star (Noirhomme-Fraiture, 2000)

- **Approches basées sur l'échantillonnage** : Birch (Zhang *et al.*, 1996), CluStream (Aggarwal *et al.*, 2003), StreamSamp (Csernel, 2008), etc.

Notre problématique



La **fouille de données d'usage du Web** (Web Usage Mining, WUM)

Ensemble de techniques basées sur la fouille de données pour analyser l'usage d'un site Web (Cooley *et al.*, 1999).

Le log Web

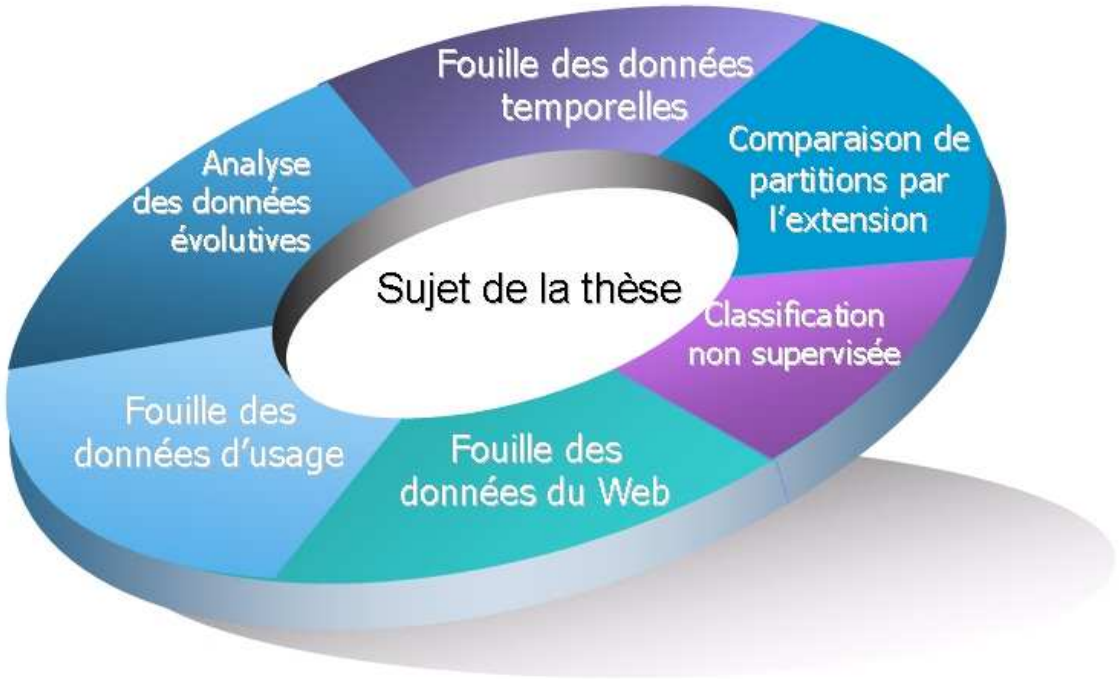
Fichier enregistré par le serveur Web contenant les traces d'usage laissées par les internautes lors des connections

- l'adresse IP du client
- la date et l'heure de la requête
- la page consultée
- le code de statut (valeur 200 en cas de réussite)
- le nombre d'octets transmis
- la page précédemment visitée
- le navigateur Web et le système d'exploitation de la machine cliente

L'unité base de notre analyse

Navigation : ensemble de requêtes (clics) proches en provenance d'un même utilisateur

Positionnement de la thèse



- **Soit** :

E : ensemble de données à classifier

$H^{(b)}$: sous-ensemble de E tel que $H^{(b)} \neq \emptyset$,

$$\forall i \neq j : H^{(i)} \cap H^{(j)} = \emptyset \text{ et } \bigcup_{b=1}^{B} H^{(b)} = E$$

Partitionnement des données par paquets (fenêtres)

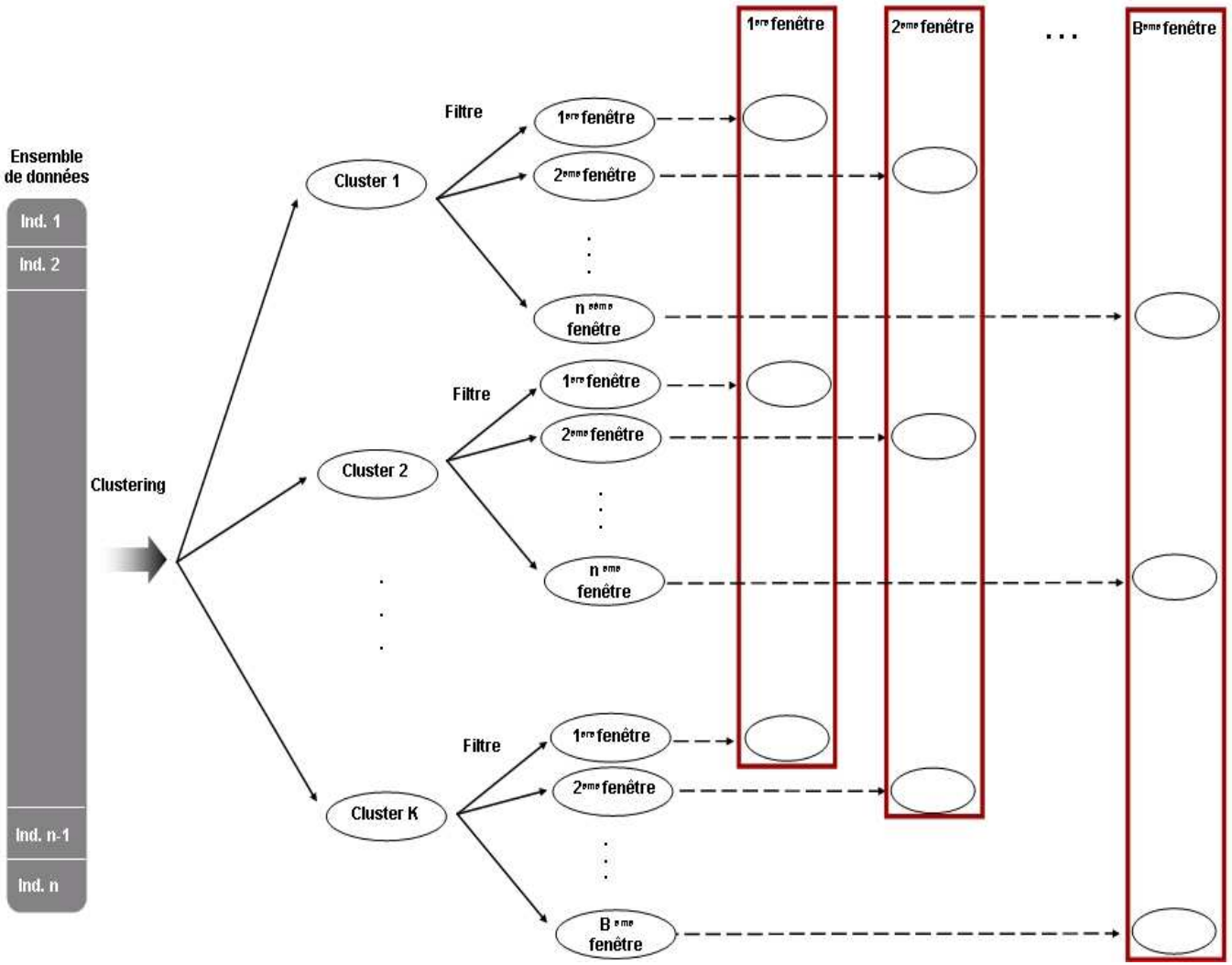
1. **Partitionnement par nombre d'effectifs constant** (fenêtre logique) on fixe le nombre d'individus qui doivent être contenus dans chaque fenêtre.

2. **Partitionnement par intervalle de temps constant** (fenêtre temporelle) on fixe un intervalle de temps durant lequel les données analysées seront enregistrées dans une fenêtre, par exemple 30 minutes, 2 heures, 1 jour, etc.

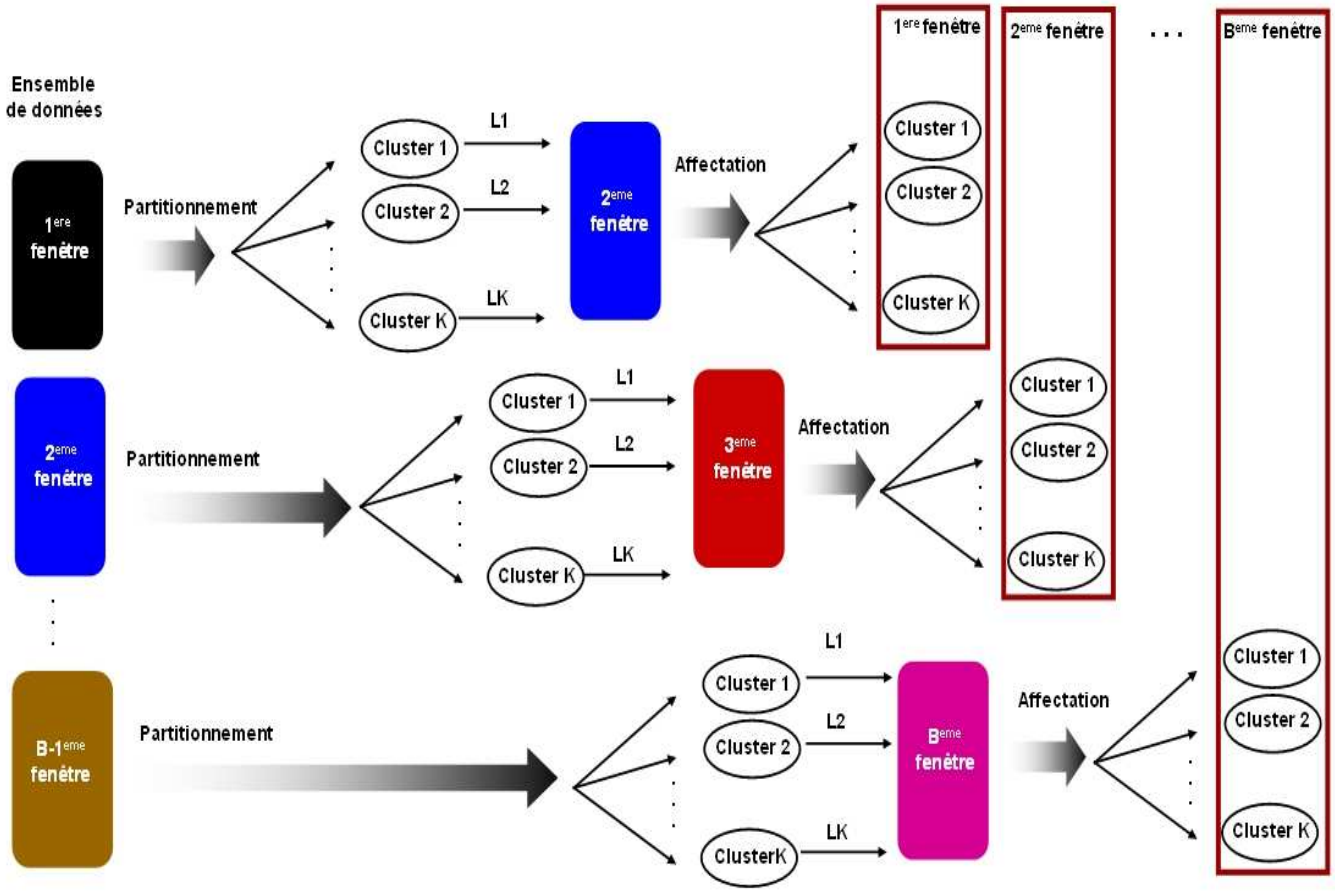
Stratégies de classification analysées

1. Classification globale (CG)
2. Classification locale précédente (CGL_1)
3. Classification locale dépendante (CGL_2)
4. Classification locale indépendante (CLI)

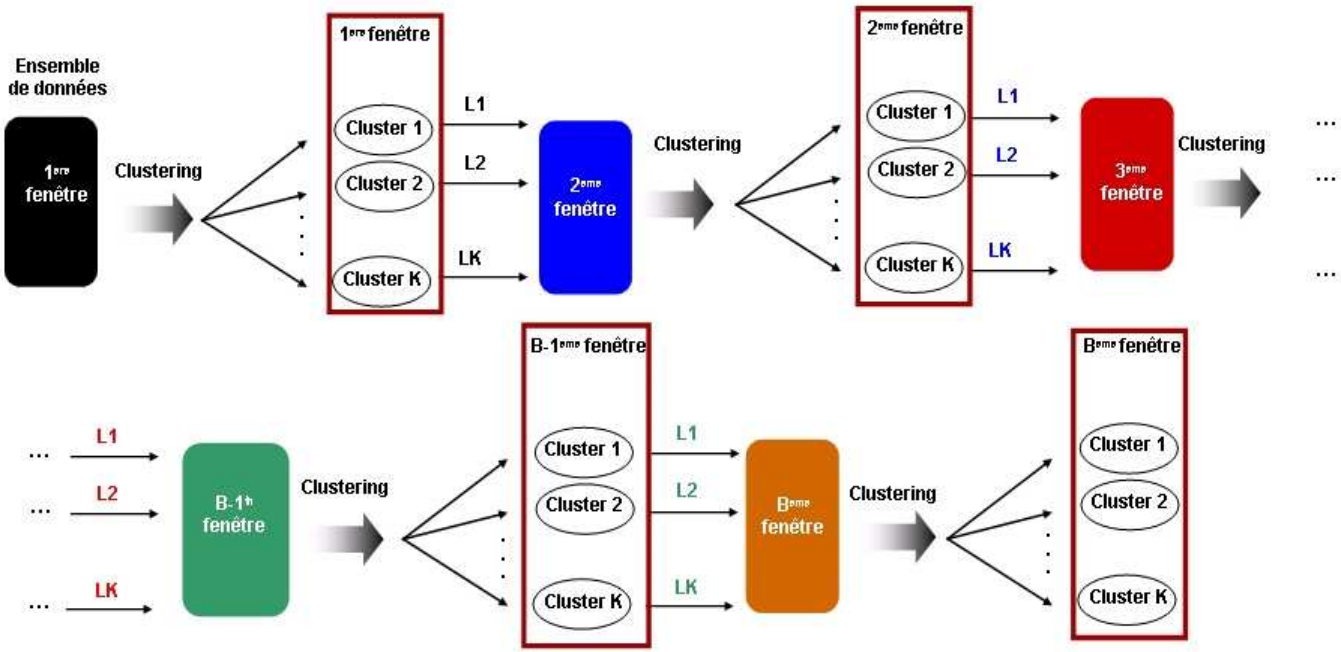
Classification globale (*CG*)



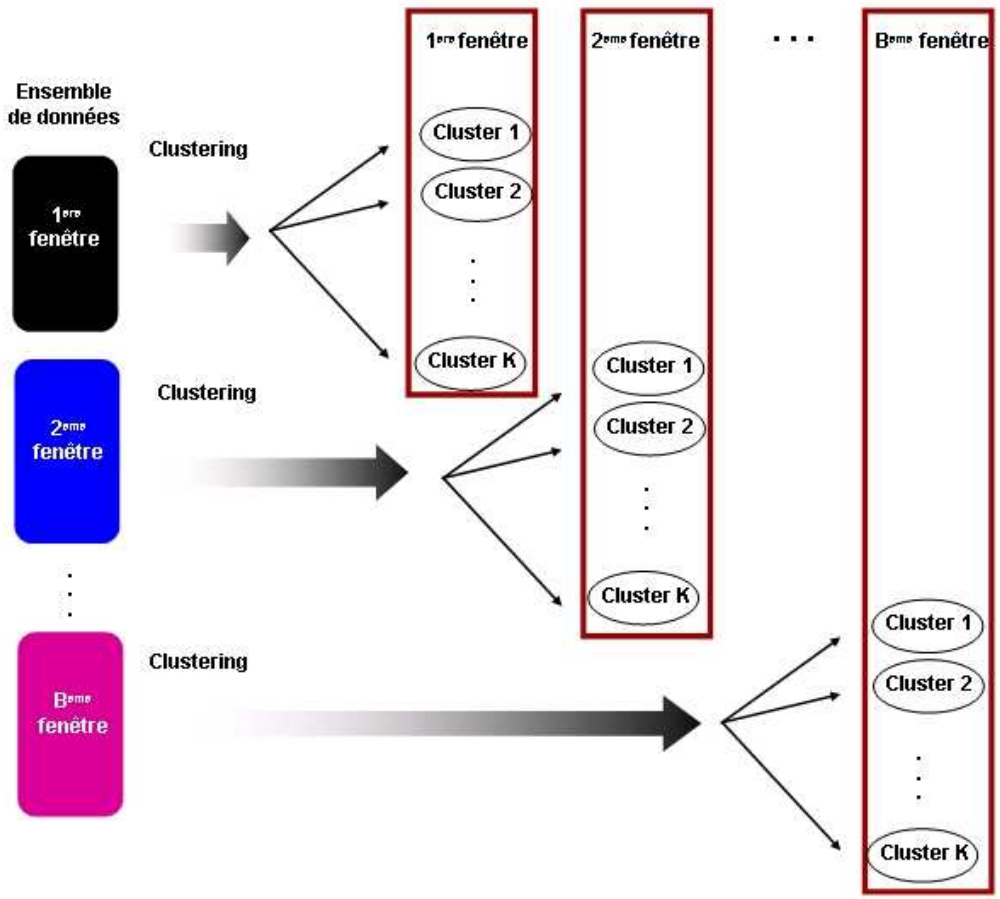
Classification locale précédente (CGL_1)



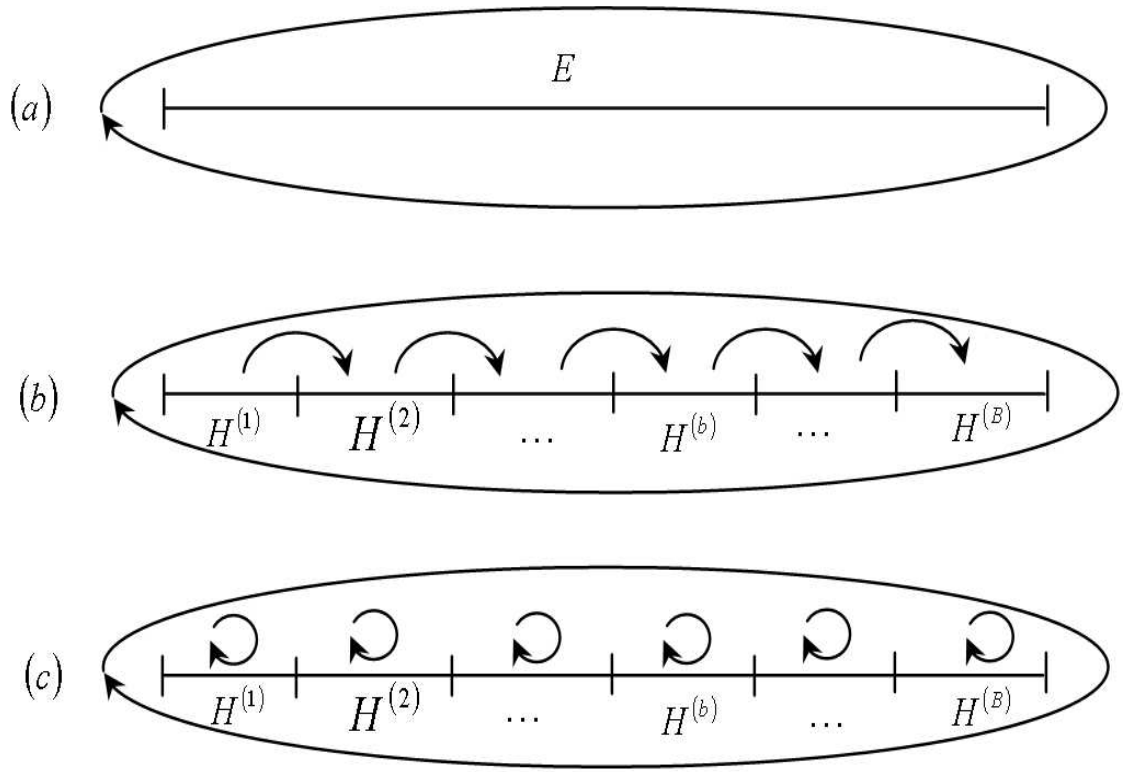
Classification locale dépendante (CGL_2)



Classification locale indépendante (*CLI*)



Comparatif des stratégies de classification



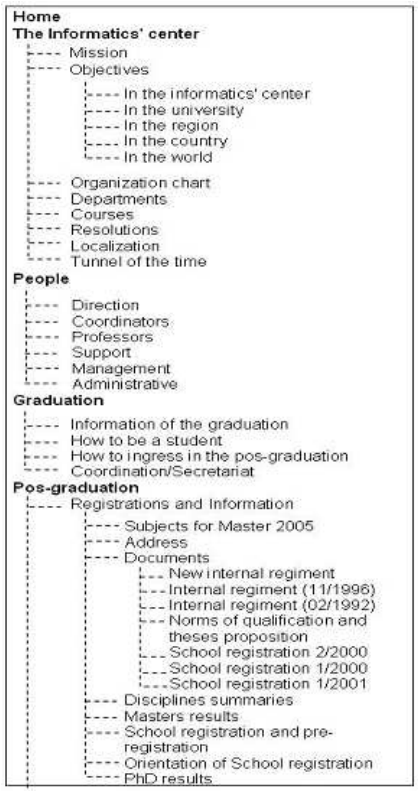
(*a*) Classification **globale** *CG* sur l'ensemble *E* [MND (Diday, 1971)]

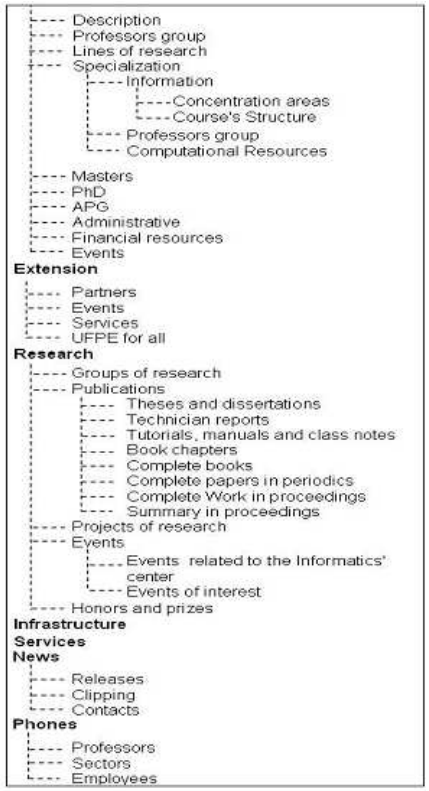
(*b*) Classification **locale précédente** CGL_1 sur les paquets $H^{(b)}$ de *E* [MNDS (Diday, 1975)]

(*c*) Classification **locale dépendante** CGL_2 sur les paquets $H^{(b)}$ de *E* [MNDSO]

Analyse d'un site Web académique [1]

Aperçu du site Web du Centre d'Informatique (CIn) de l'UFPE :





1. http://www.cin.ufpe.br/

Site Web du Centre d'Informatique (CIn) de l'UFPE

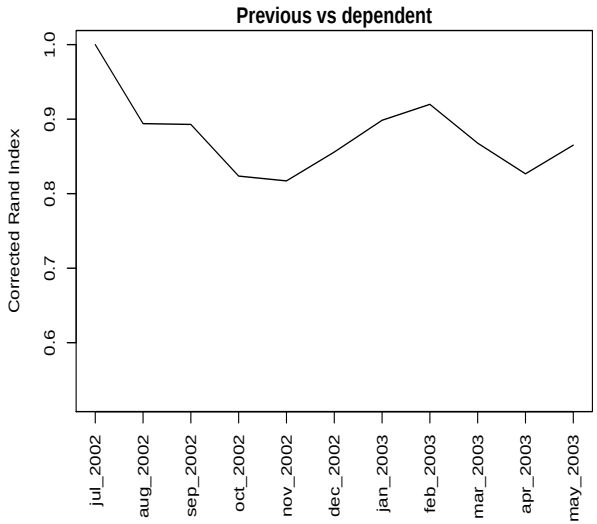
- Composition :
  - pages statiques (pages personnelles, pages de support de cours, etc.)
  - 90 pages pages dynamiques gérées par des *servlets* en Java
- Période analysée :
  - du 01 juillet 2002 à 00 :10 :25 jusqu'au 15 mai 2003 à 13 :22 :27

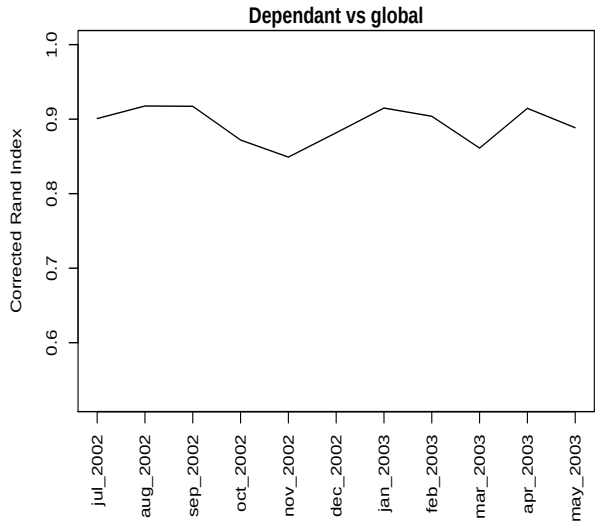
Paramètres de l'analyse

distance :	Euclidienne
nombre de clusters K :	10
nombre d'initialisations :	50
type de fenêtre :	temporelle de taille égale à 1 mois
données d'entrée :	tableau *navigations x variables statistiques*
total d'individus :	138 536 navigations

Valeurs de RC entre les résultats des stratégies de classification deux-à-deux :

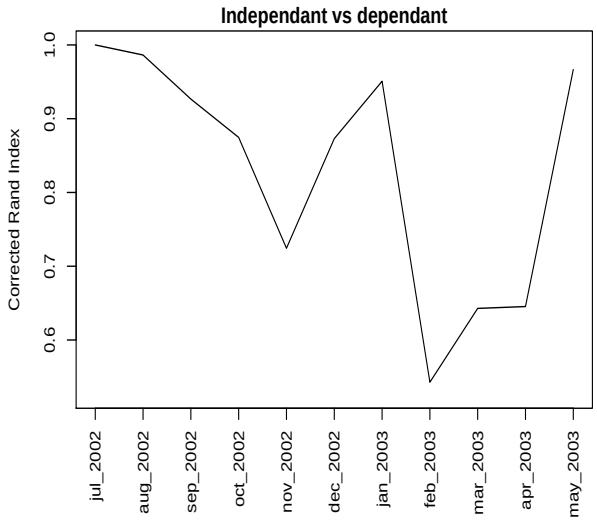
MNDS sur $H^{(b)}$
versus
MNDSO sur $H^{(b)}$

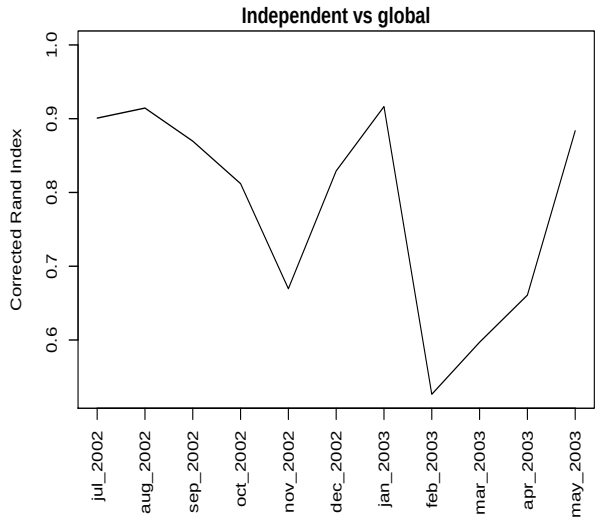




MND sur E
versus
MNDSO sur $H^{(b)}$

MND sur $H^{(b)}$ (CLI)
versus
MNDSO sur $H^{(b)}$

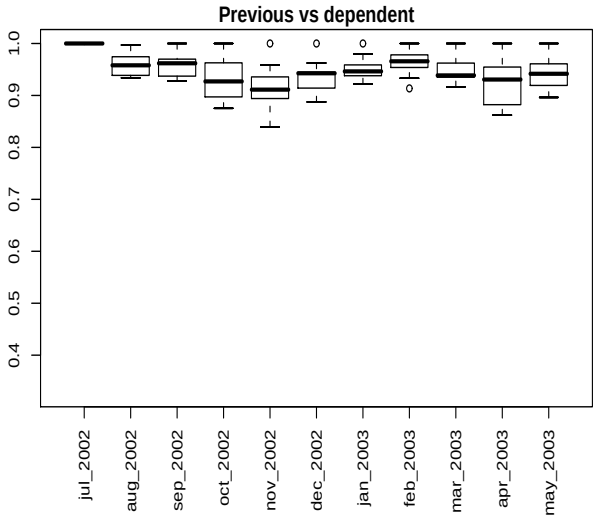


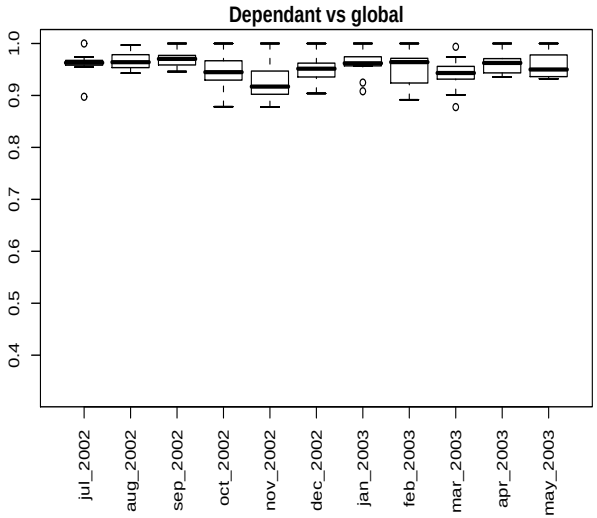


MND sur $H^{(b)}$ (CLI)
versus
MND sur E

Valeurs de F-measure entre les résultats des stratégies de classification deux-à-deux :

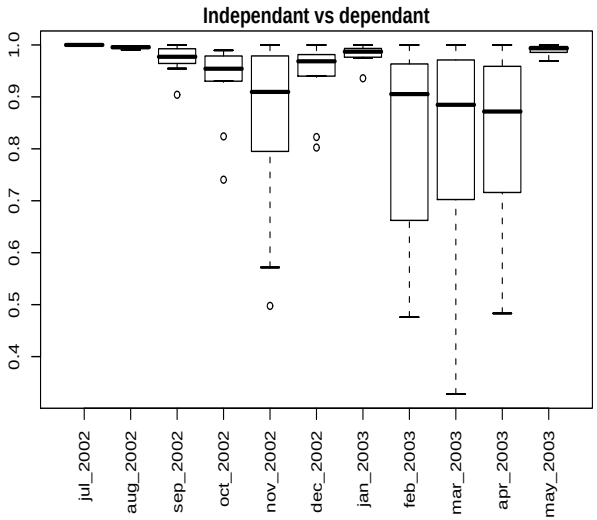
MNDS sur $H^{(b)}$
versus
MNDSO sur $H^{(b)}$

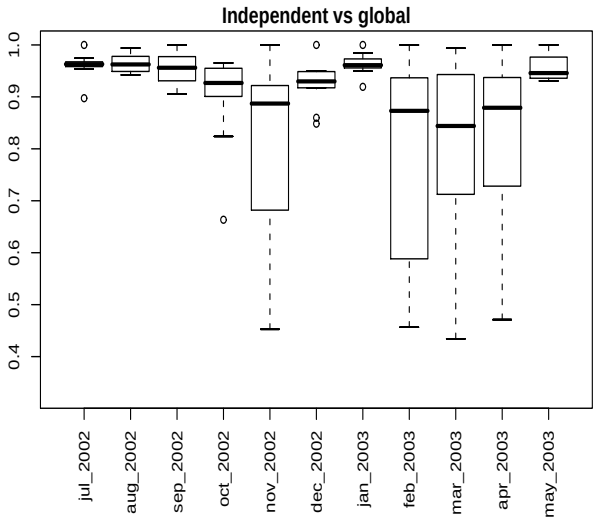




MND sur E
versus
MNDSO sur $H^{(b)}$

MND sur $H^{(b)}$ (CLI)
versus
MNDSO sur $H^{(b)}$



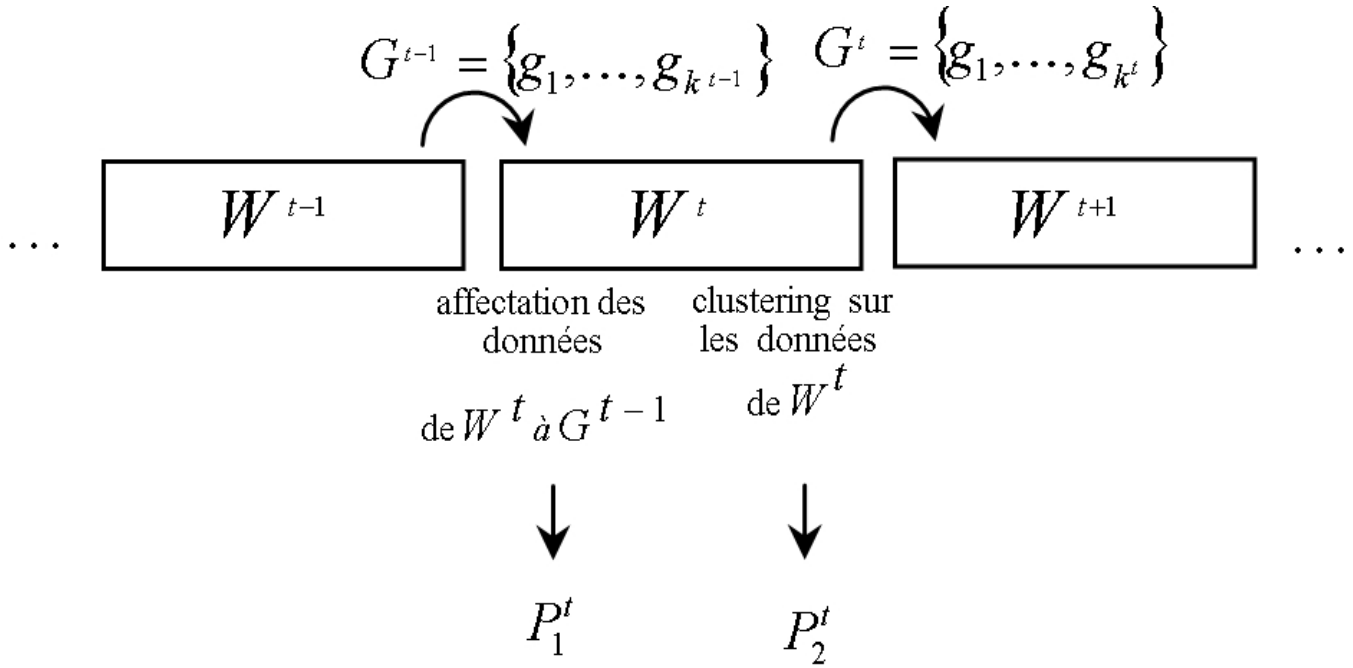


MND sur $H^{(b)}$ (CLI)
versus
MND sur E

Conclusion

- Ces résultats confirment de façon expérimentale les résultats théoriques démontrés dans le manuscrit :
  - $\Rightarrow$ les méthodes MND, MNDS et MNDSO convergent vers les mêmes solutions
- La stratégie de classification locale indépendante ($CLI \Rightarrow$ MND sur $H^{(b)}$) est la plus adaptée à la détection des changements entre deux fenêtres de temps

Schéma de l'approche de classification pour la détection et le suivi des changements :



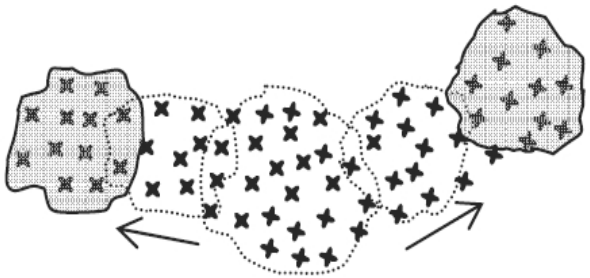
Stratégie de détection de changements

P_1^t : données courantes segmentées selon la structure classificatoire d'un passé recent
P_2^t : données courantes segmentées sans prise en compte du passé

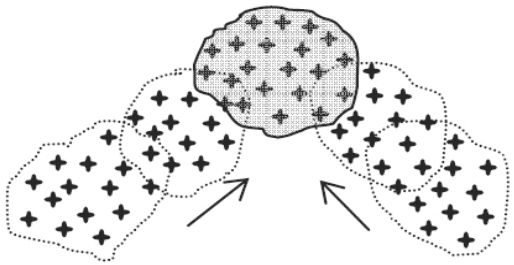
$$P_1^t = P_2^t \, ?$$

Interprétation des changements

- **Scission ou division d'un cluster** : un certain nombre d'individus appartenant à un cluster de la partition P_1^t migre vers d'autres clusters dans la partition P_2^t

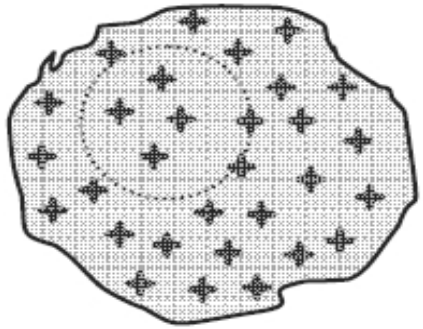


- **Fusion ou absorption de deux ou plusieurs clusters** : les individus de différents clusters de la partition P_1^t migrent vers un même cluster de la partition P_2^t



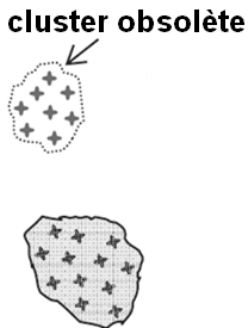
Interprétation des changements

- **Pas de changement externe** : le cluster est dit survivant dans la fenêtre de temps suivante



Interprétation des changements

- **Disparition d'un cluster** : un cluster de la partition P_1^t n'est plus présent dans la partition P_2^t



- **Apparition d'un cluster** : un cluster inexistant dans la partition P_1^t émerge dans la partition P_2^t

Détermination du nombre de classes

- Question largement exploitée dans la littérature d'apprentissage automatique (Milligan & Cooper, 1985) (Dubes, 1987) (Halkidi & Vazirgiannis, 2001)

- *"Une classification ne peut pas être varie ou fausse, ni probable ou improbable, mais seulement profitable ou non profitable"* (Williams & Lance, 1965)

- Découverte du vrai nombre de classes ?
  - Nombre de clusters acceptable/utile dans un but précis

- Segmenter les individus de manière à faire ressortir des préférences d'usage (pages Web, produits de commerce, etc.)

Les indices de détermination du nombre de classes comparés dans notre étude :

1. la pseudo-statistique de Calinski et Harabasz (CH) [ou G1* sur R]
2. l'indice de Baker et Hubert (BH) [ou G2* sur R]
3. l'indice de Hubert et Levin (HL) [ou G3* sur R]
4. l'indice Silhouette (S)
5. l'indice de Davies et Bouldin (DB)
6. l'indice de Krzanowski et Lai (KL)
7. l'indice de Hartigan (H)
8. la statistique Gap (G)
9. stratégie de détermination du nombre de clusters basée sur les dérivées première et seconde (D)

*Package clusterSim de **R** :
`http://cran.r-project.org/web/packages/clusterSim`

Stratégie de détermination du nombre de classes

1. Appliquer le SOM (Kohonen, 1995) avec une grille contenant 200 neurones initialisés par une ACP (Elemento, 1999)
2. Appliquer une CAH avec le critère de Ward sur les prototypes (neurones) finaux de la SOM (Murtagh, 1995)
3. Tracer le graphique des gains d'inertie intra-classe obtenus à chaque itération de l'algorithme CAH
4. Le nombre de classes à retenir sera obtenu par :
   - L'application des differents indices de determination du nombre de classes [1-8]
   - Le coude dans la décroissance des valeurs du gain d'inertie intra-classe, repéré à l'aide des dérivées (différences) premières et secondes [D] (Lebart *et al.*, 1995)

Expérimentation sur des données artificielles

Méthodologie de génération de données

Un algorithme de base de génération des données artificielles d'usage

```
1  Pour c dans l'intervalle[1,K] {
2    Pour chaque individu i de la classe c {
3      x_i ← {0,...,0}
4      nbClic ← random (nbClicMIN, nbClicMAX)
5      Pour y dans l'intervalle[1,nbClic] {
6        z ← random(0,1)
7        Pour j dans l'intervalle[1,p] {
8          Si (z <= cumul(p_c^j)) {
9            x_i^j ← x_i^j + 1
10           Sortir de la boucle la plus interne
11         }
12       }
13     }
14   }
15 }
```

Exemple :

$c_1 = (p_1^1, p_1^2, p_1^3) = (0.3, 0.4, 0.3)$

$cumul(c_1) =$

(p_1^1)	$(p_1^1 + p_1^2)$	$(p_1^1 + p_1^2 + p_1^3)$
0.3	0.7	1.0

$nbClic = 10$

$z = random(0,1)$

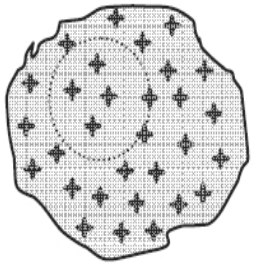
$P(z \leq cumul(p_c^j))$

Expérimentation sur des données artificielles

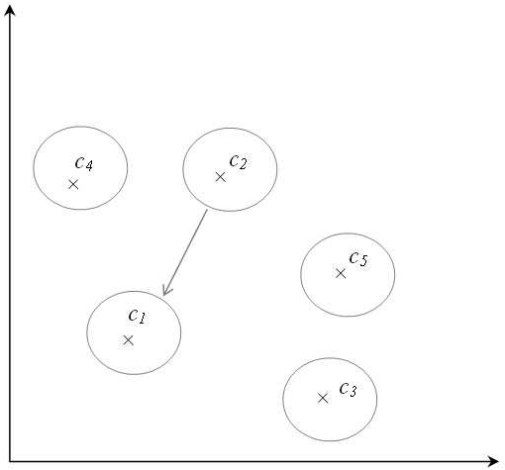
Méthodologie de génération de données

Trois algorithmes spécialisés dans la simulation de changements sur des données artificielles

1. Changement lié à l'effectif des classes
   - Rétrécissement d'une classe (facteur β_1)
   - Grossissement d'une classe (facteur β_2)



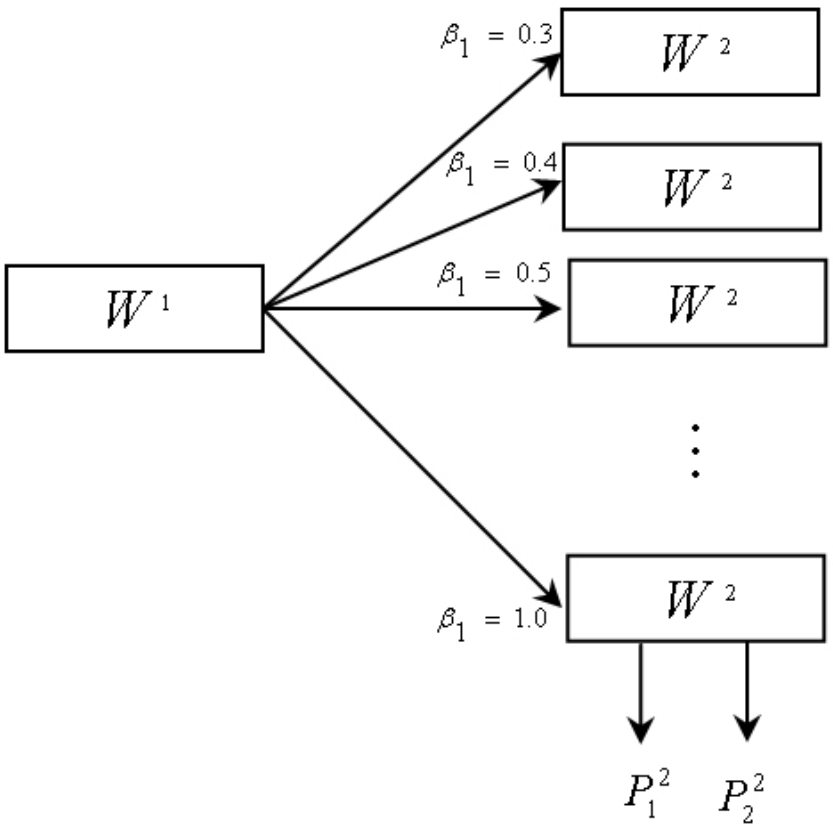
2. Changement lié à la position des classes (facteur β_3) dans l'espace des données



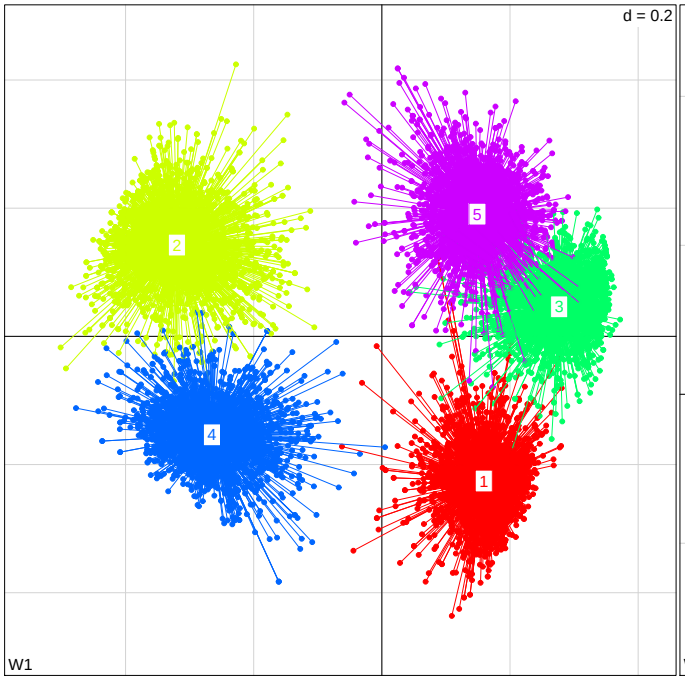
Expérimentation sur des données artificielles

Méthodologie de génération de données

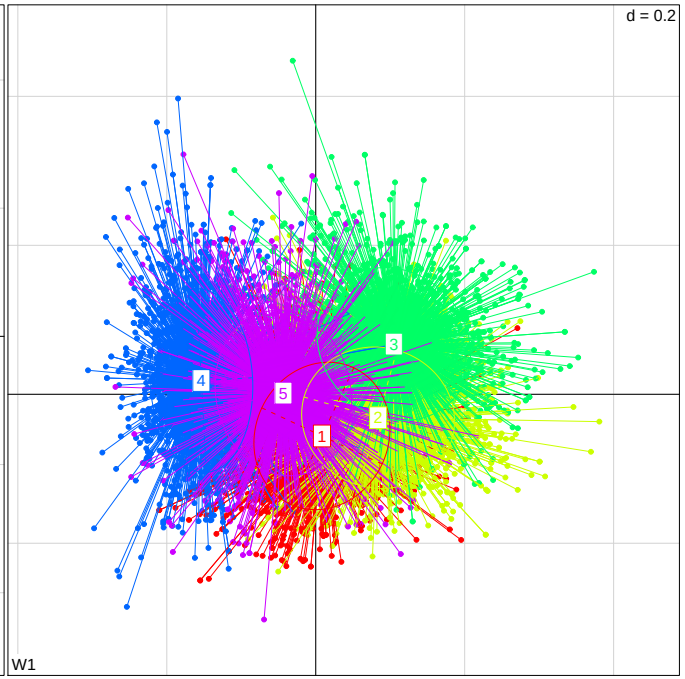
Génération de la fenêtre W^2 à partir de W^1 (cas de rétrécissement) :



Projection des classes artificielles contenues dans W^1

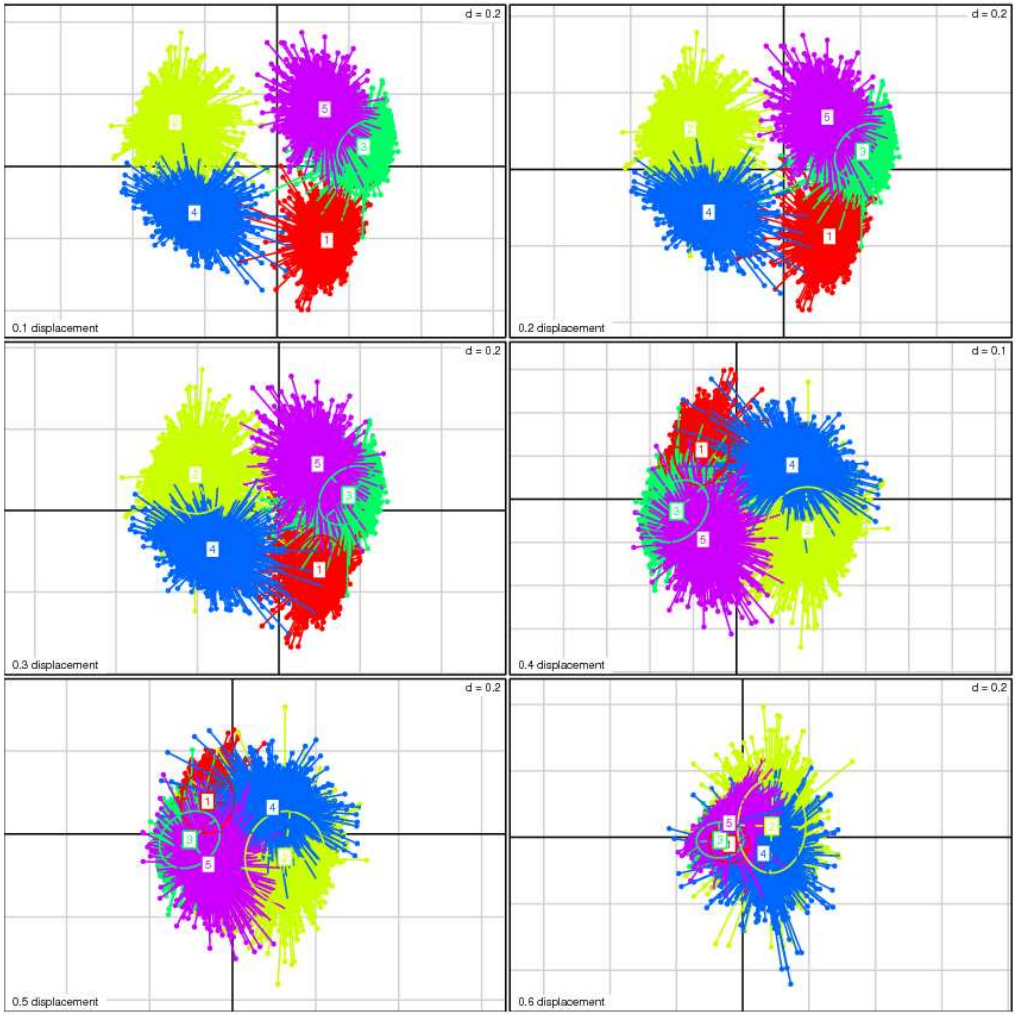


classes bien séparées



classes recouvrantes

Simulation de l'effet de rapprochement des classes



Expérimentation sur des données artificielles

Le tableau suivant discrimine le nombre de clusters suggéré par chaque indice pour chaque fenêtre de données analysée.

Fenêtre	CH	BH	HL	KL	DB	H	G	S	D
1	5	5	10	4	5	5	5	5	5
2	5	30	5	4	5	5	8	5	5
3	5	30	6	4	5	5	6	5	5
4	2	30	14	4	30	5	6	4	5
5	2	30	9	4	30	5	5	5	5
6	2	30	9	9	29	5	8	7	5
7	10	30	9	9	30	2	2	10	4
%correct	42.85	14.28	14.28	0.00	42.85	85.71	28.57	57.14	85.71

Meilleure performance présentée par :

- **l'indice de Hartigan**
- **la stratégie basée sur des dérivées**

Expérimentation sur des données réelles

- Centre d'informatique de l'UFPE (Recife, Brésil) [2]
- Structure dynamique contenant 90 pages organisées sur 10 thèmes : *The Informatics' center*, *People*, *Graduation*, *Pos-graduation*, *Extension*, *Research*, *Infrastructure*, *Services*, *News* et *Phones*.

Paramètres

méthode de classification :	SOM (Kohonen, 1995) initialisée par une ACP suivie d'une CAH avec le critère de Ward
distance :	Euclidienne
type de fenêtre :	temporelle de taille égale à 1 mois
données d'entrée :	tableau *navigations x thèmes de pages*
nombre de thèmes de pages :	10
filtre appliqué :	navigations ayant un minimum de 10 clics
total d'individus :	28 220 navigations

2. `http://www.cin.ufpe.br/`

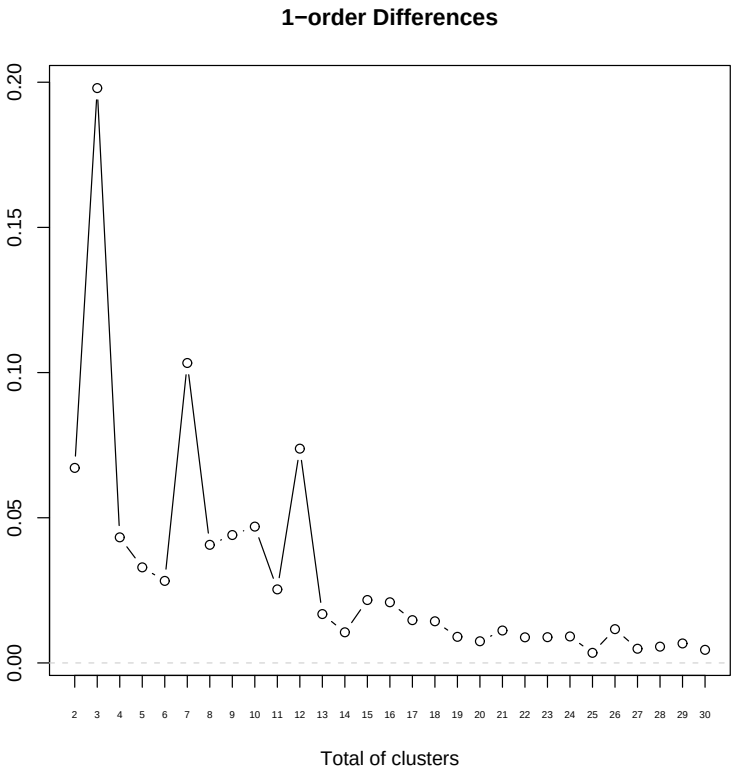
Expérimentation sur des données réelles

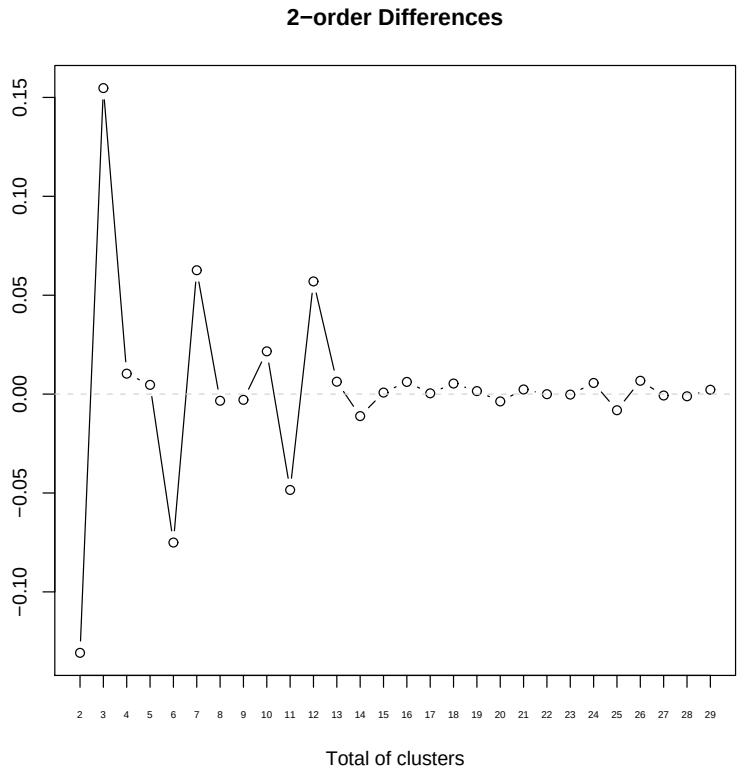
Nombre de classes suggéré par les différents indices pour le jeu de données d'usage du CIn-UFPE :

Fenêtre	CH	BH	HL	KL	DB	H	G	S	D
1	30	30	6	2	29	4	3	3	3
2	3	30	2	2	30	5	5	21	3
3	2	2	6	7	2	3	2	2	5
4	2	2	11	3	2	4	2	2	4
5	30	30	12	9	26	2	2	29	3
6	29	30	8	8	30	3	2	30	3
7	2	2	13	24	2	10	3	2	2
8	30	30	8	8	30	2	2	30	4
9	30	25	11	2	30	3	3	25	3
10	30	30	2	4	25	5	3	30	3
11	30	30	7	4	28	2	2	30	5

Résultat obtenu par le critère des dérivées pour la première fenêtre analysée

Les différences premières et secondes des inerties intra-classes relatives des partitions contenant de 2 à 30 classes :





Conclusion de l'étude de cas sur les indices de détermination du nombre de classes

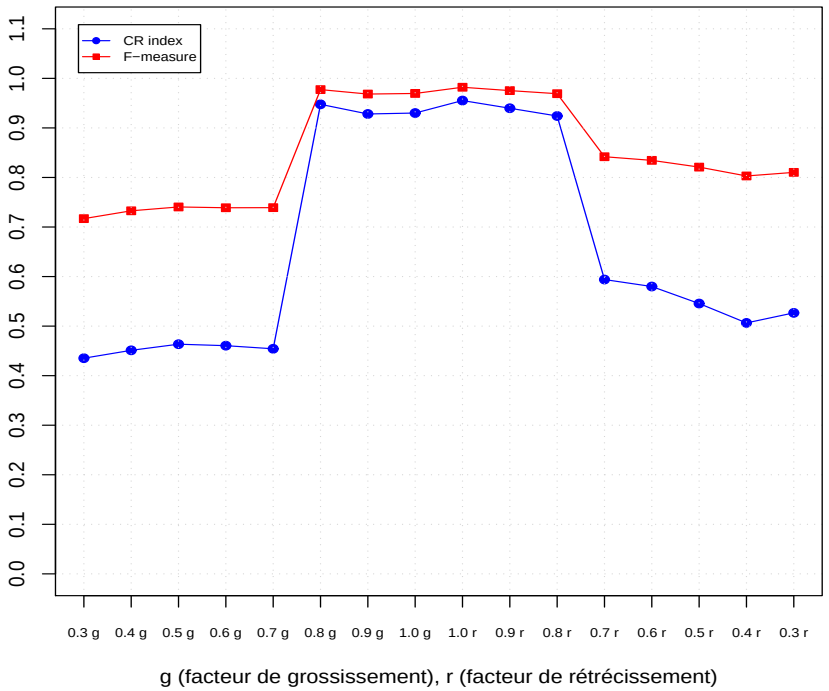
- Les différents indices ne sont pas toujours d'accord sur le nombre de clusters à retenir sur un même jeu de données
  - le choix du nombre de clusters s'avère être une tâche délicate

- Après plusieurs analyses sur différents jeux de données :
  - La stratégie basée sur les dérivées semble être très efficace

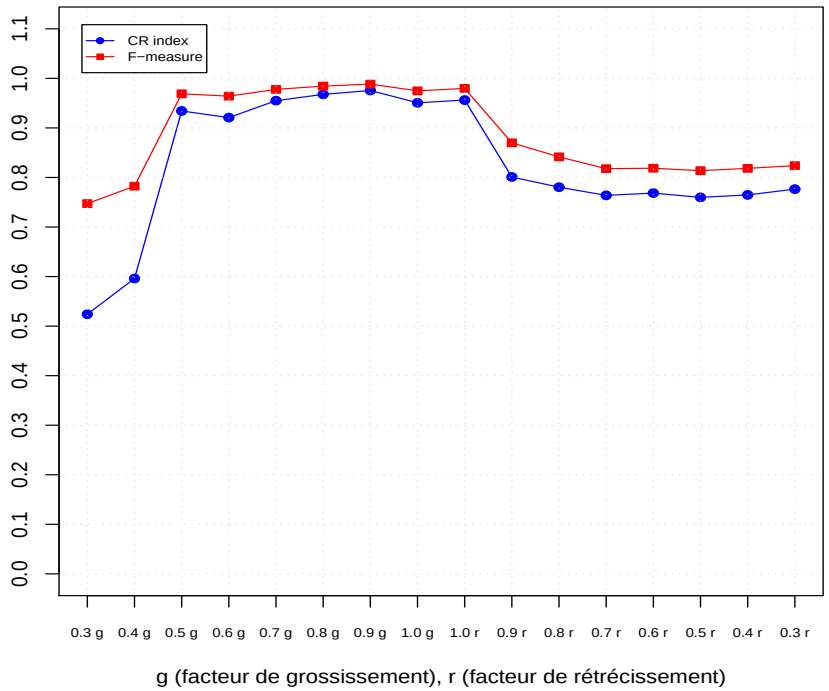
Application de l'approche sur des données artificielles

Grossissement et rétrécissement de la classe cible

Valeurs de l'indice corrigé de Rand et de la F-measure :



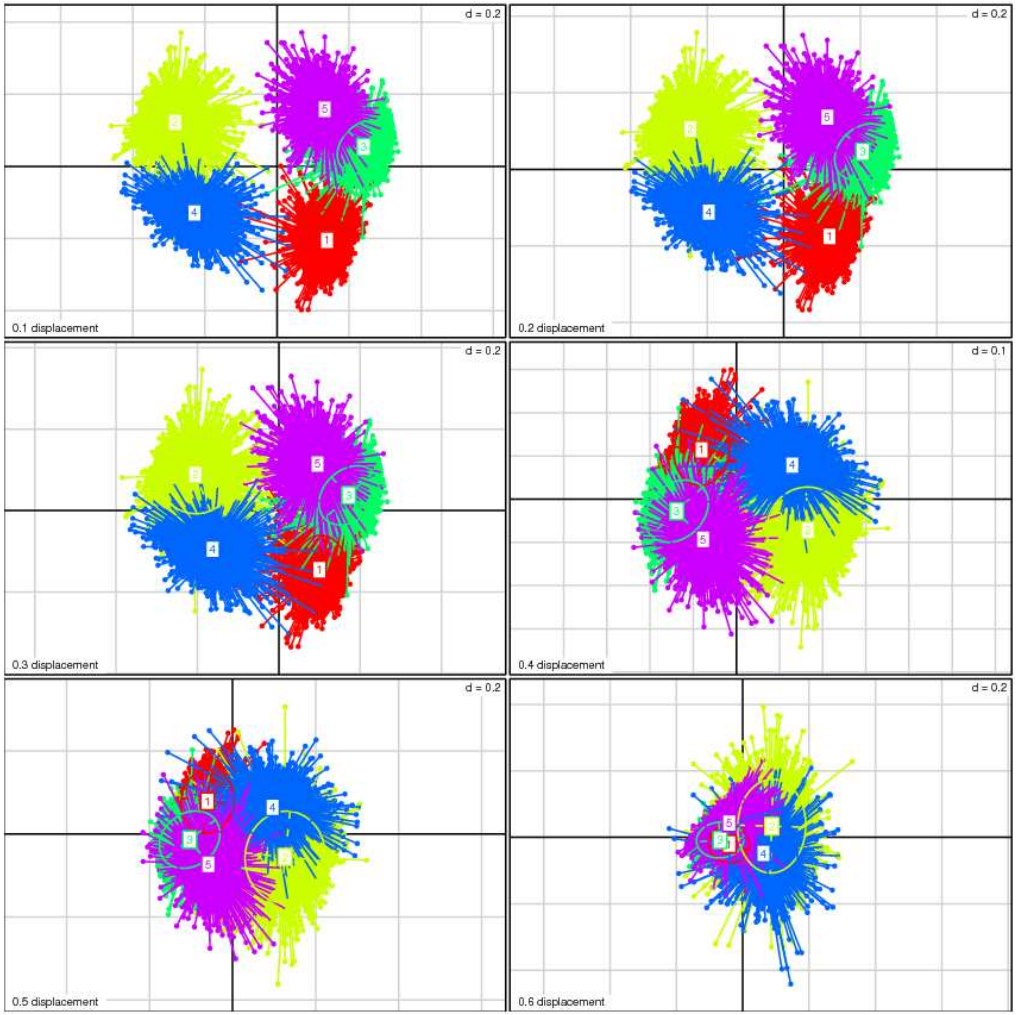
classes bien séparées



classes recouvrantes

Application de l'approche sur des données artificielles

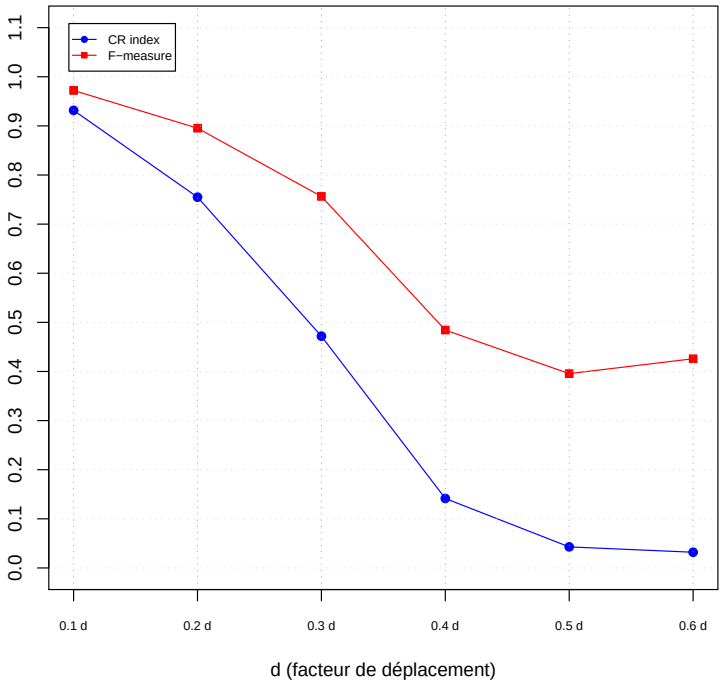
Simulation de l'effet de déplacement des classes



Application de l'approche sur des données artificielles

Simulation de l'effet de déplacement des classes

Valeurs de l'indice corrigé de Rand et de la F-measure :

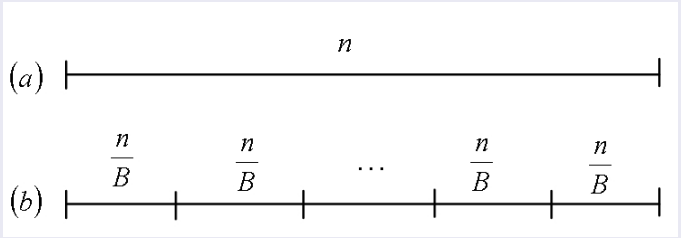


Conclusion de l'étude de cas sur les données artificielles

- L'approche proposée se montre très efficace pour la détection des changements
- Les deux indices utilisés pour la comparaison de partitions (Rand corrigé et F-measure) se montrent capables de repérer les changements dans le temps
  - L'indice corrigé de Rand étant plus sensible aux changements que la F-mesure

Caractérisation et avantages de l'approche de classification

1. Stratégie du type *diviser pour régner*
2. Réduction de la complexité de la méthode de classification. Ex. : $O(n^2) \rightarrow BO((\frac{n}{B})^2)$



3. Indépendance de la méthode de classification
4. Processus de détection des changements basé sur l'extension
5. Application des algorithmes non-incrémentaux dans un processus incrémental
6. Utilisation d'un passé récent (mémoire)
7. Résumé des données au cours du temps

Schéma de la base de données

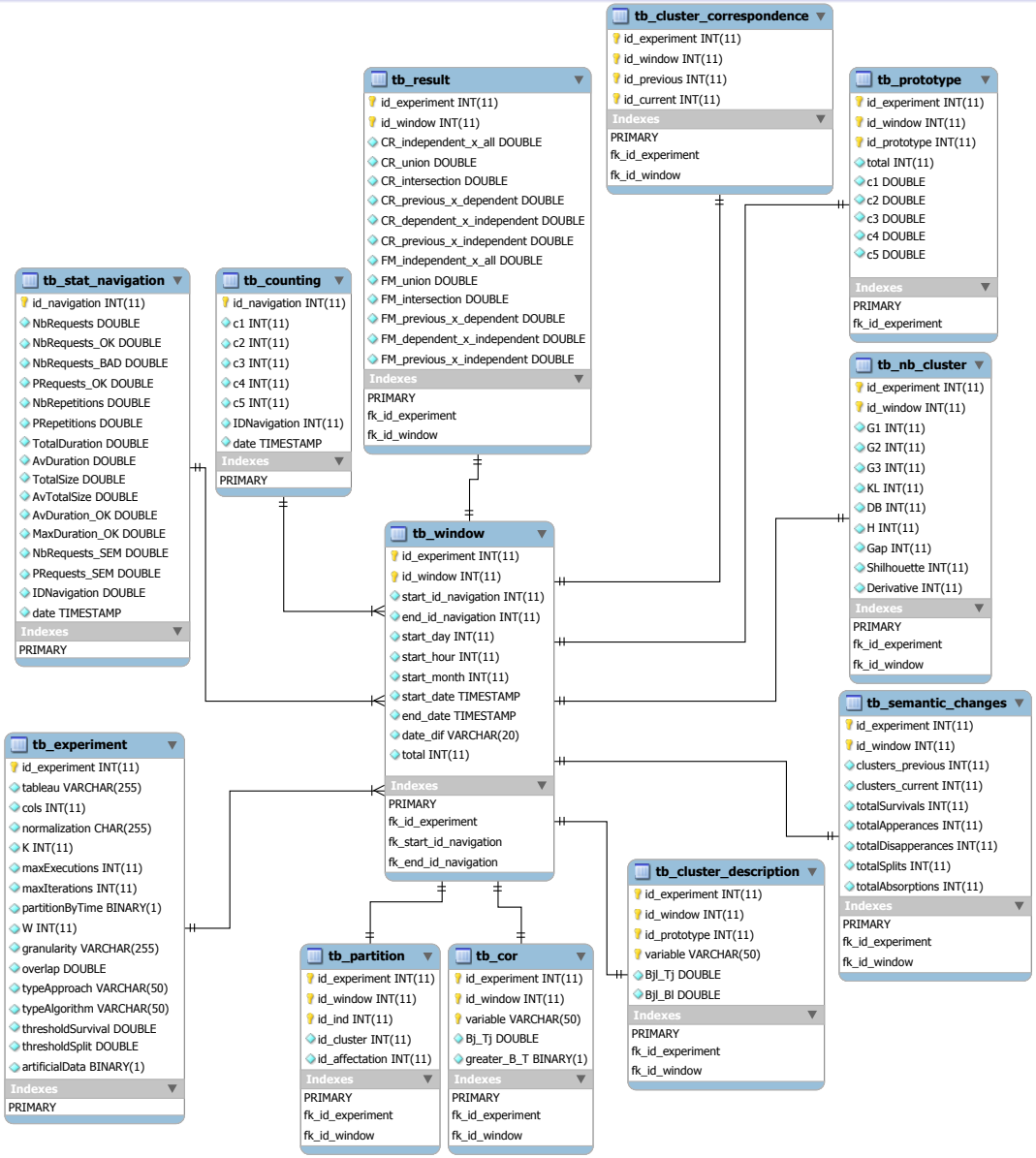
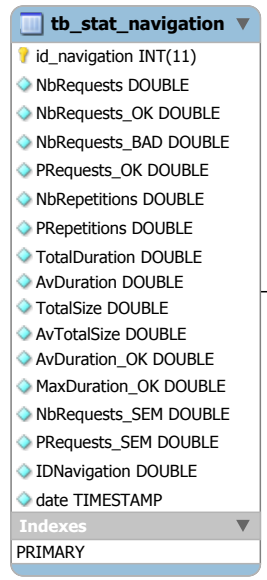
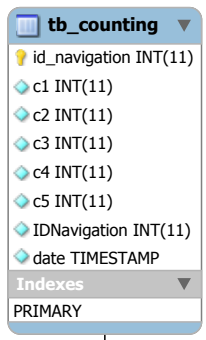
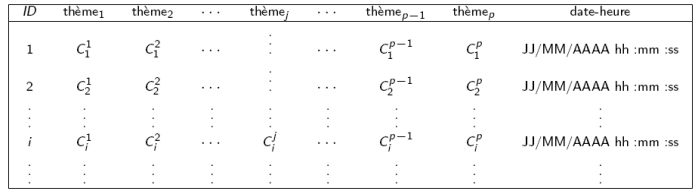


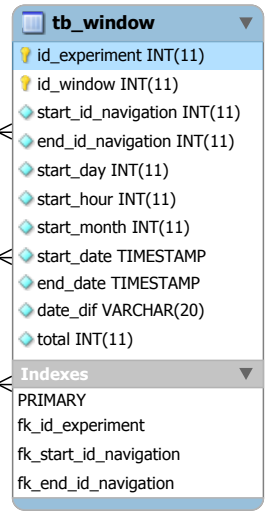
Schéma de la base de données

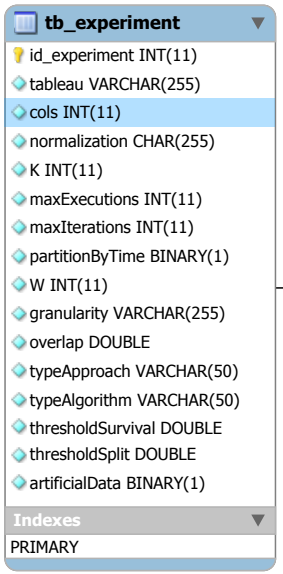
N°	Variable	Signification
1	NbRequests	Nombre de clics effectués durant la navigation
2	NbRequests_OK	Nombre de requêtes réussies (statut = 200) dans la navigation
3	NbRequests_Bad	Nombre de requêtes échouées (statut ≠ 200) dans la navigation
4	PRequests_OK	Pourcentage de requêtes réussies (= *NbRequests_OK / NbRequests*)
5	NbRepetitions	Nombre de requêtes répétées dans la navigation
6	PRepetitions	Pourcentage de répétitions (= *NbRepetitions / NbRequests*)
7	DureeTotale	Durée totale de la navigation (en secondes)
8	MDuree	Moyenne de la durée des requêtes (= *DureeTotale / NbRequests*)
9	MDuree_OK	Moyenne de la durée des requêtes réussies (= *DureeTotale_OK / NbRequests_OK*)
10	NbRequests_Sem	Nombre de requêtes liées aux pages de la structure sémantique du site
11	PRequests_Sem	Pourcentage de requêtes liées aux pages de la structure sémantique du site (= *NbRequests_Sem / NbRequests*)
12	TotalSize	Total d'octets transférés durant la navigation
13	MSize	Moyenne d'octets transférés durant la navigation (= *TotalSize / NbRequests_OK*)
14	DureeMax_OK	Durée maximale parmi les durées des requêtes réussies











- tb_result
- tb_partition
- tb_cluster_correspondance
- tb_prototype
- tb_nb_cluster
- tb_semantic_changes
- tb_cluster_description
- tb_cor

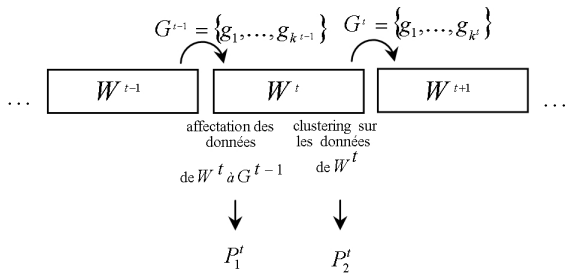
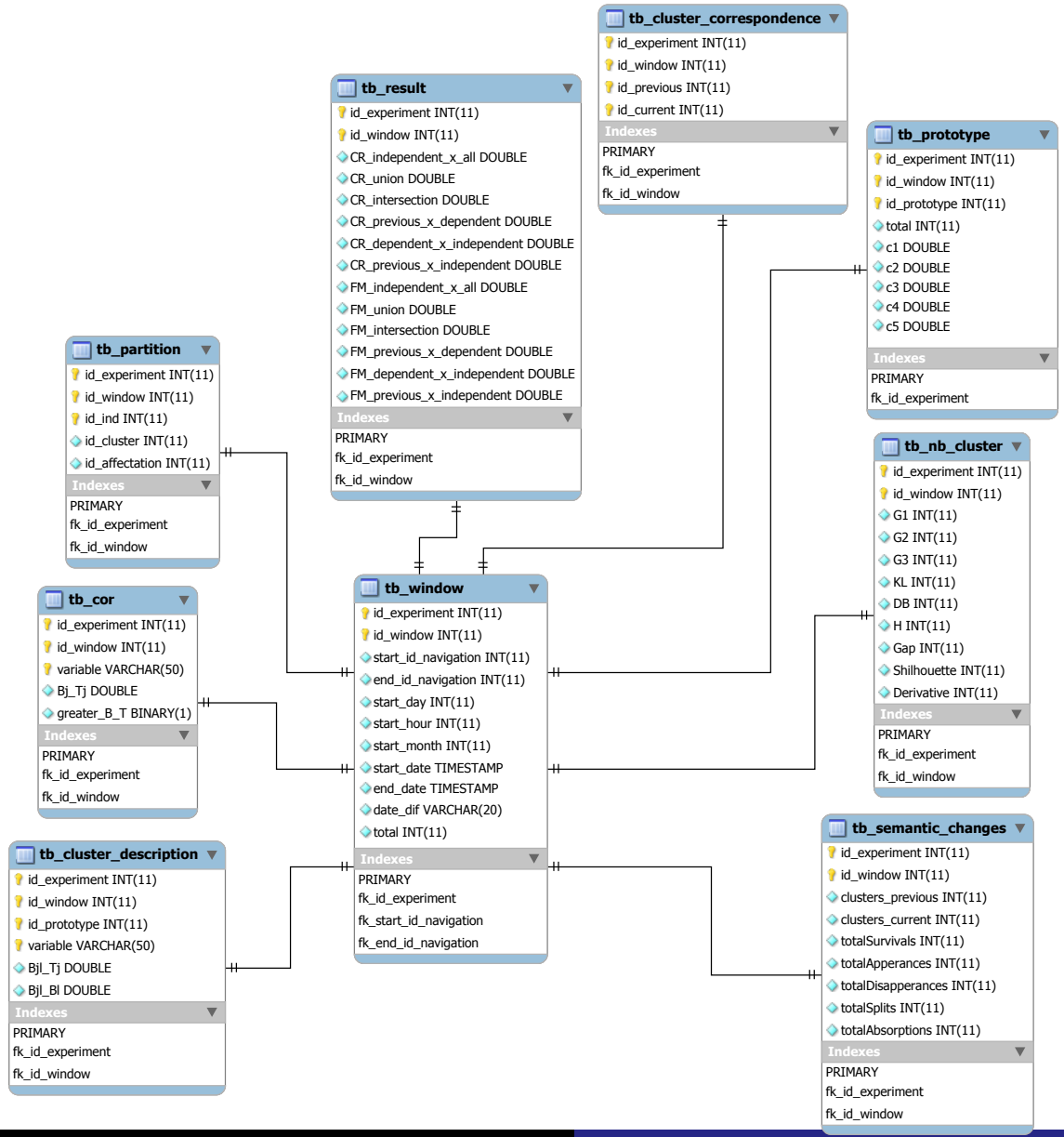


Schéma de la base de données



Analyse d'un jeu de données issu du *marketing*

- Suivi des achats d'un panel de consommateurs mis à disposition dans le cadre du SLDS 2009 [3]
  - 10 068 clients
  - 2 marchés de biens de consommation
  - 6 marques de produits
- Période analysée :
  - 14 mois (du 09 juillet 2007 jusqu'au 08 septembre 2008)
- Champs descriptifs :

Champ	Signification
ident	identification client
date	premier jour de la semaine de l'achat
marché	marché concerné (marche_1, marche_2)
marque	marque du produit acheté (A, B, C, D, E, F)
valeur	valeur des achats

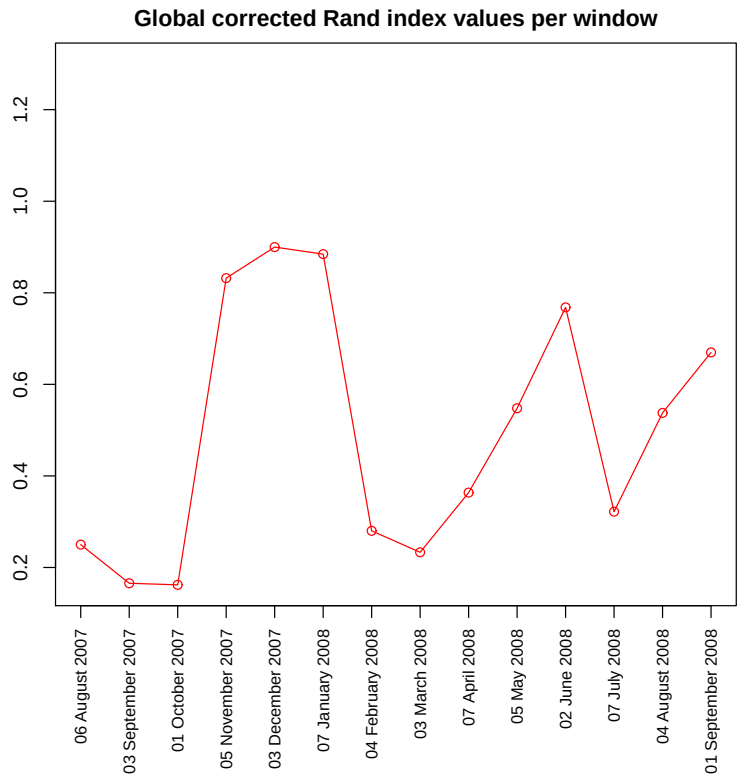
3. `http://www.ceremade.dauphine.fr/SLDS2009`

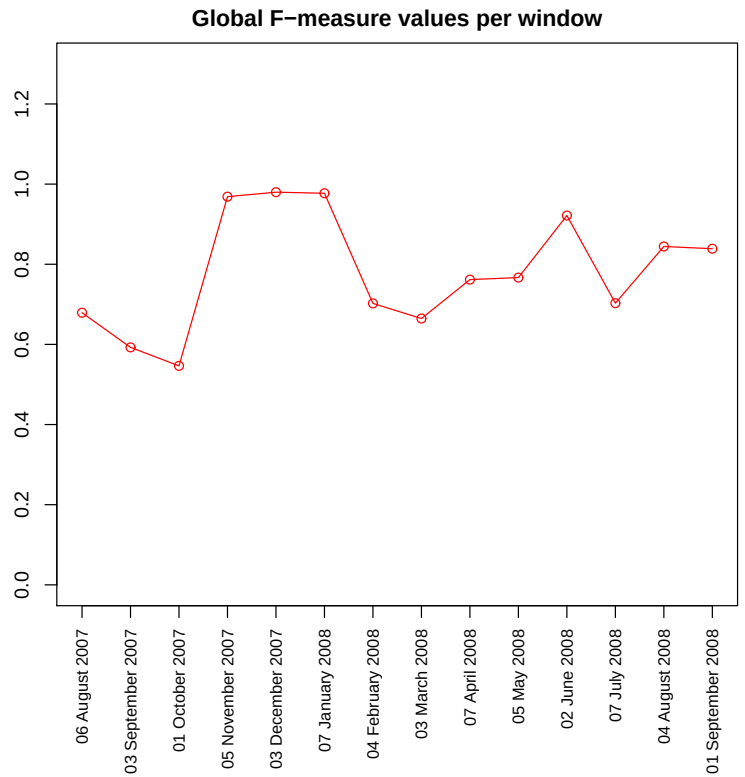
Analyse d'un jeu de données issu du *marketing*

Paramètres

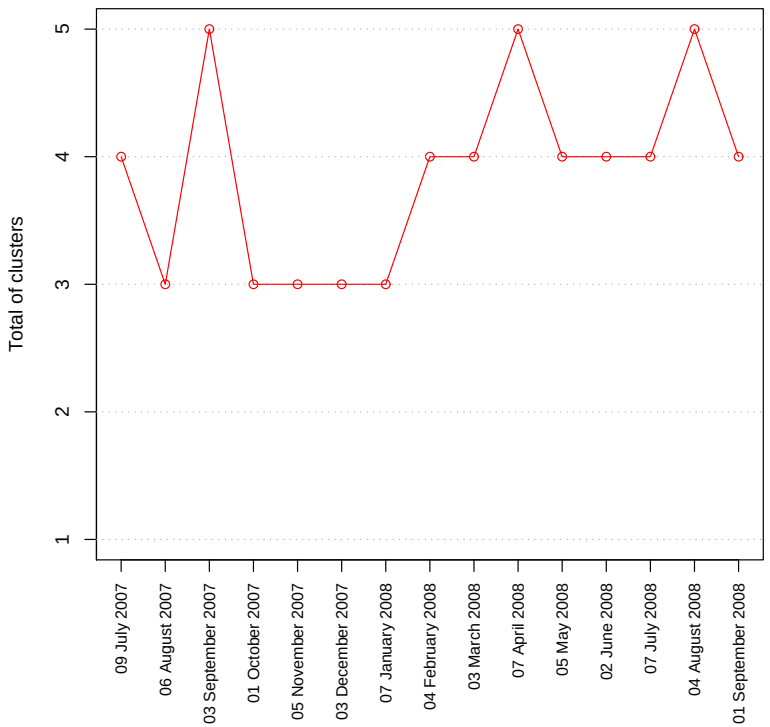
méthode de classification :	SOM (Kohonen, 1995) initialisée par une ACP suivie d'une CAH avec le critère de Ward, puis du critère basé sur les dérivées
distance :	Euclidienne
nombre de dimensions :	6 (marques de produit)
type de fenêtre :	temporelle de taille égale à 1 mois
données d'entrée :	tableau croisé *client x marques de produit*
total d'individus :	262 215

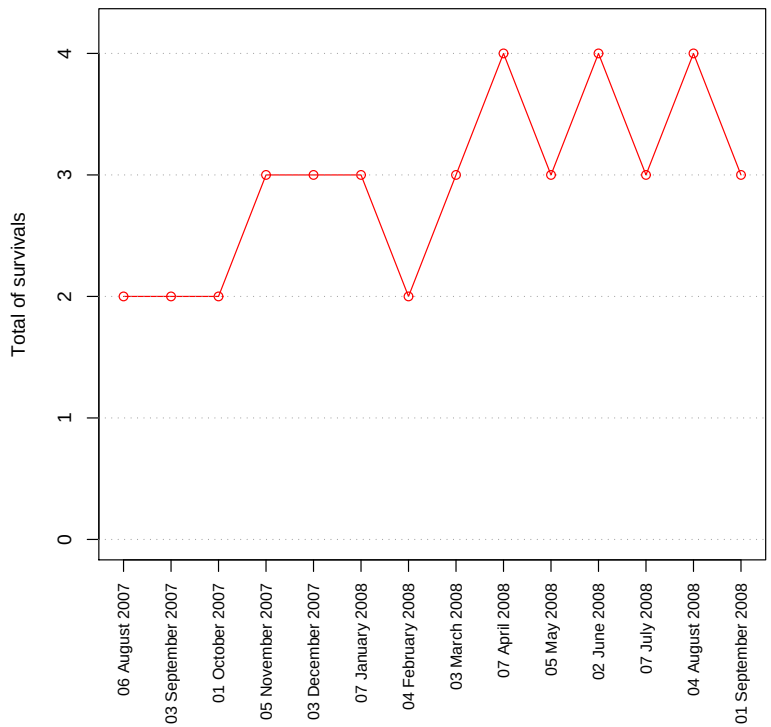
- Valeurs obtenues par l'indice de Rand corrigé (à gauche) et par la F-mesure (à droite)
- Quelques périodes d'instabilité : mois de *Octobre 2007*, *Mars 2008* et *Juillet 2008*.





- D'*Octobre 2007* jusqu'à *Janvier 2008* : le nombre total de clusters est égal à 3.
- Ces mois correspondent à une période de grande stabilité selon montrent les valeurs élevées de l'**indice de Rand** et de la **F-measure**





Interprétation statistique des classes par variable (Celeux *et al.*, 1989)

Le pouvoir discriminant de la variable j par rapport à la partition P

$$COR(j) = \frac{B^j}{T^j}$$

La contribution relative de la variable j à l'inertie inter-classe de la partition P

$$CRT(j) = \frac{B^j}{B} \text{ avec } \sum_{j=1}^{p} CRT(j) = 1$$

Le pouvoir discriminant de la variable j par rapport à la classe C_l

$$COR(j,l) = \frac{B_l^j}{T^j} \text{ avec } COR(j) = \sum_{l=1}^{K} COR(j,l)$$

La contribution relative de la variable j à l'interie inter-classe de C_l

$$CTR(j,l) = \frac{B_l^j}{B_l}$$

Trois premières variables les plus significatives des clusters de comportement issus du mois d'*octobre 2007*.

id cluster	variable	$\frac{B_l^j}{T^j}$	$\frac{B_l^j}{B_l}$	$\frac{B^j}{T^j}$	$> \frac{B}{T}$
1	F	2.23e-006	0.44	3.56e-005	oui
1	B	2.19e-006	0.43	3.11e-005	oui
1	A	2.18e-007	0.04	1.78e-006	non
2	F	3.32e-005	0.93	3.56e-005	oui
2	A	1.31e-006	0.03	1.78e-006	non
2	E	4.24e-007	0.01	1.04e-006	non
3	B	2.88e-005	0.93	3.11e-005	oui
3	D	8.01e-007	0.02	1.38e-006	non
3	E	4.69e-007	0.01	1.04e-006	non

Ces profils d'achat peuvent être décrits comme suit :

- **Cluster 1** : groupe des clients ayant une préférence d'achat majoritaire pour les marques F et B en premier plan et la marque A en deuxième plan
- **Cluster 2** : groupe des clients ayant une préférence d'achat pour la marque F en premier plan et les marques A et E en deuxième plan
- **Cluster 3** : groupe des clients ayant une préférence d'achat pour la marque B en premier plan et les marques D et E en deuxième plan

Conclusion de l'analyse de données du *marketing*

- Possibilité d'application de l'approche de détection et suivi des changements sur d'autres domaines

- Pour des stratégies de CRM, le profil d'un client pourrait être interprété comme suit :
  - Si le client reste dans un même cluster au cours du temps : client *fidèle* aux marques de produit
  - Si le client change considérablement de cluster au cours du temps : client *zappeur* entre les marques de produit

Conclusion de l'analyse de données du *marketing*

- Possibilité d'application de l'approche de détection et suivi des changements sur d'autres domaines

- Pour des stratégies de CRM, le profil d'un client pourrait être interprété comme suit :
  - Si le client reste dans un même cluster au cours du temps : client *fidèle* aux marques de produit
  - Si le client change considérablement de cluster au cours du temps : client *zappeur* entre les marques de produit

Apports

- ✔ Deux modélisations des données d'usage du Web
  1. Caractérisation du mode de navigation de l'internaute
  2. Caractérisation du centre d'intérêt de l'internaute
- ✔ Une méthodologie de génération de données d'usage et de simulation de changements
- ✔ Analyse théorique et expérimentale des méthodes de classification MND, MNDS et MNDSO.

Apports

- ✔ Une approche de classification permettant la détection et le suivi des changements sur des données évolutives
  - Opère sur de fenêtres non recouvrantes de taille préfixée
  - Indépendante de la méthode de classification
  - Applique des indices de comparaison de partition basés sur l'extension
  - Intègre l'interprétation des changements repérés

- ✔ Applicabilité de l'approche à d'autres domaines

  Ex. : Surveillance de matériels (tâche WP3.2, projet MIDAS)

- ✔ Outil d'analyse mis en ligne :
  - http://atwueda.gforge.inria.fr/

Perspectives

↪ La mise en œuvre de techniques permettant de déterminer de façon adaptative la taille de la fenêtre à adopter

↪ L'application d'autres indices de validation externe permettant de ressortir les divergences entre deux partitions (par exemple, l'indice dérivé de Mac Nemar (Youness & Saporta, 2004), etc.)

↪ Représentation des prototypes des classes par des intervalles avec l'ADS (Ciampi & Lechevallier, 2000) (Hardy, 2009)

↪ Amélioration de la partie d'interprétation des changements
- Modélisation à l'aide des techniques de l'aide à la décision

Merci de votre attention !

Questions ?

Publications majeures

- DA SILVA, Alzennyr, LECHEVALLIER, Yves, ROSSI, Fabrice, DE CARVALHO, Francisco. Clustering Dynamic Web Usage Data. In Nadia Nedjah, Luiza Mourelle and Janusz Kacprzyk (Editors), Springer Berlin/Heidelberg, Series : Studies in Computational Intelligence, *Innovative Applications in Data Mining*, Vol. 169, pages 71-82, ISBN : 978-3-540-88044-8, 2009.

- DA SILVA, Alzennyr, LECHEVALLIER, Yves, DE CARVALHO, Francisco. Monitoring Data Changes through a Clustering Approach. In *International Federation of Classification Societies* (IFCS 2009). Dresden, March 13-18, 2009.

- DA SILVA, Alzennyr. Diverses approches permettant l'introduction du temps dans la fouille de données d'usage du Web. Editeurs : Chantal Reynaud et Gilles Venturini. *Numéro spécial sur la fouille du Web de la Revue des Nouvelles Technologies de l'Information* (RNTI W-1), pages 35-55, ISBN 978.2.85428.793.6, cépaduès éditions, 2007.

- DA SILVA, Alzennyr. Analyzing the Evolution of Web Usage Data. In *Special issue on Data Stream Analysis of MODULAD* (Monde des Utilisateurs de L'Analyse de Données), numéro 36, pages 75-84, May, 2007.

- DA SILVA, Alzennyr, LECHEVALLIER, Yves, ROSSI, Fabrice, DE CARVALHO, Francisco. Construction and Analysis of Evolving Data Summaries : an Application on Web Usage Data. In Luiza Mourelle, Nadia Nedjah and Janusz Kacprzyk editors, *VII IEEE International Conference on Intelligent Systems Design and Applications* (ISDA 2007), Pages 377-380, ISBN : 978-0-7695-2976-9, IEEE Computer Society, Rio de Janeiro, Brazil, 22-24 October, 2007.

- DA SILVA, Alzennyr, DE CARVALHO, Francisco, LECHEVALLIER, Yves, TROUSSE, Brigitte. Mining Web Usage Data for Discovering Navigation Clusters. In : *XI IEEE Symposium on Computers and Communications* (ISCC 2006), pages 910-915, ISBN ĨSSN :1530-1346, 0-7695-2588-1, Mining Web Usage Data for Discovering Navigation Clusters, IEEE Computer Society, Pula-Cagliari, Italy, 2006.

- DA SILVA, Alzennyr, DE CARVALHO, Francisco, LECHEVALLIER, TROUSSE, Brigitte. Characterizing Visitor Groups from Web Data Streams. In : *IEEE International Conference on Granular Computing* (GrC 2006), Atlanta, USA, 2006, pages 389-392, ISBN : 1-4244-0134-8, IEEE Computer Society, 2006.

Comparaison de deux partitions

Tableau de contingence entre les partitions P_1^t et P_2^t :

clusters de P_1^t \ clusters de P_2^t	C_1	$\cdots$	C_j	$\cdots$	C_k	
C_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	$n_{1.}$
$\vdots$						
C_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ik}	$n_{i.}$
$\vdots$						
C_m	n_{m1}	$\cdots$	n_{mj}	$\cdots$	n_{mk}	$n_{m.}$
	$n_{.1}$		$n_{.j}$		$n_{.k}$	$n_{..} = n$

Indices basés sur l'extension : F-measure

Rappel et Précision

$$rappel(C_i, C_j) = \frac{n_{ij}}{\sum_{j=1}^{k} n_{ij}} \; ; \; precision(C_i, C_j) = \frac{n_{ij}}{\sum_{i=1}^{m} n_{ij}}$$

F-measure entre deux clusters

$$Fmeasure(C_i, C_j) = \frac{2precision(C_i, C_j)rappel(C_i, C_j)}{precision(C_i, C_j) + rappel(C_i, C_j)}$$

F-measure entre deux partitions

$$F(P_1^t, P_2^t) = \frac{1}{n} \sum_{i=1}^{m} n_{i.} \max_{j=1,\ldots,k} Fmeasure(C_i, C_j)$$

Indices basés sur l'extension : indice corrigé de Rand

Indice corrigé de Rand (RC) entre deux partitions

$$RC(P_1^t, P_2^t) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{m}\binom{n_{i.}}{2}\sum_{j=1}^{k}\binom{n_{.j}}{2}}{\frac{1}{2}[\sum_{i=1}^{m}\binom{n_{i.}}{2} + \sum_{j=1}^{k}\binom{n_{.j}}{2}] - \binom{n}{2}^{-1}\sum_{i=1}^{m}\binom{n_{i.}}{2}\sum_{j=1}^{k}\binom{n_{.j}}{2}}$$

Variables descriptives des navigations

Le nombre d'accès aux pages Web (page hit count) a long temps été utilisé comme indicateur majeur des préférences des internautes (Yan *et al.*, 1996).

N°	Variable	Signification
1	NbRequests	Nombre de clics effectués durant la navigation
2	NbRequests_OK	Nombre de requêtes réussies (statut = 200) dans la navigation
3	NbRequests_Bad	Nombre de requêtes échouées (statut $\neq$ 200) dans la navigation
4	PRequests_OK	Pourcentage de requêtes réussies (= *NbRequests_OK* / *NbRequests*)
5	NbRepetitions	Nombre de requêtes répétées dans la navigation
6	PRepetitions	Pourcentage de répétitions (= *NbRepetitions* / *NbRequests*)
7	DureeTotale	Durée totale de la navigation (en secondes)
8	MDuree	Moyenne de la durée des requêtes (= *DureeTotale* / *NbRequests*)
9	MDuree_OK	Moyenne de la durée des requêtes réussies (= *DureeTotale_OK* / *NbRequests_OK*)
10	NbRequests_Sem	Nombre de requêtes liées aux pages de la structure sémantique du site
11	PRequests_Sem	Pourcentage de requêtes liées aux pages de la structure sémantique du site (= *NbRequests_Sem* / *NbRequests*)
12	TotalSize	Total d'octets transférés durant la navigation
13	MSize	Moyenne d'octets transférés durant la navigation (= *TotalSize* / *NbRequests_OK*)
14	DureeMax_OK	Durée maximale parmi les durées des requêtes réussies

Tableau de comptage des clics sur des thèmes de pages

Représentation des données d'usage en tant que flux de navigations :

ID	thème_1	thème_2	$\cdots$	thème_j	$\cdots$	thème_{p-1}	thème_p	date-heure
1	C_1^1	C_1^2	$\cdots$	$\vdots$	$\cdots$	C_1^{p-1}	C_1^p	JJ/MM/AAAA hh :mm :ss
2	C_2^1	C_2^2	$\cdots$	$\vdots$	$\cdots$	C_2^{p-1}	C_2^p	JJ/MM/AAAA hh :mm :ss
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	C_i^1	C_i^2	$\cdots$	C_i^j	$\cdots$	C_i^{p-1}	C_i^p	JJ/MM/AAAA hh :mm :ss
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Expérimentation sur des données artificielles

Méthodologie de génération de données

Paramètre	Valeurs utilisées lors des expérimentations
p	10
nbWindow	2
nbClicMIN	10
nbClicMAX	50
K	5
totalIndividuals	10 000
c'	1 (changement d'effectifs) et 2 (déplacement)
β_1 (facteur de rétrécissement)	$\{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
β_2 (facteur de grossissement)	$\{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
β_3 (facteur de déplacement)	$\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$

Deux cas de figures :

1. Classes bien séparées
2. Classes recouvrantes

References I


Aggarwal, Charu C. 2005.
On change diagnosis in evolving data streams.
Ieee transactions on knowledge and data engineering, **17**(5), 587–600.


Aggarwal, Charu C., Han, Jiawei, Wang, Jianyong, & Yu, Philip S. 2003.
A framework for clustering evolving data streams.
Pages 81–92 of : Vldb '2003 : Proceedings of the 29th international conference on very large data bases.
VLDB Endowment.

References II

Baron, Steffan, & Spiliopoulou, Myra. 2001.
Monitoring change in mining results.
Pages 51–60 of : Kambayashi, Yahiko, Winiwarter, Werner, & Arikawa, Masatoshi (eds), *Data warehousing and knowledge discovery (dawak)*.
Lecture Notes in Computer Science, vol. 2114.
Springer.

Bartolini, Ilaria, Ciaccia, Paolo, Ntoutsi, Irene, Patella, Marco, & Theodoridis, Yannis. 2004.
A unified and flexible framework for comparing simple and complex patterns.
Pages 496–499 of : Pkdd '04 : Proceedings of the 8th european conference on principles and practice of knowledge discovery in databases.

References III

New York, NY, USA : Springer-Verlag New York, Inc.

Celeux, G., Diday, E., Govaert, G., Lechevallier, Y., & Ralambondrainy, H. 1989.
Classification automatique des données.
Dunod, Paris.

Chen, Xiaodong, & Petrounias, Ilias. 1999.
Mining temporal features in association rules.
Pkdd'99 : Proceedings of the third european conference on principles of data mining and knowledge discovery, 295–300.

Ciampi, A., & Lechevallier, Y. 2000.
Clustering large, multi-level data sets : An approach based on kohonen self organizing maps.
Principles of data mining and knowledge discovery, springer berlin / heidelberg, **1910/2000**, 161–177.

References IV

Cooley, Robert, Mobasher, Bamshad, & Srivastava, Jaidep. 1999.
Data preparation for mining world wide web browsing patterns.

Journal of knowledge and information systems, **1**(1), 5–32.

Corvaisier, Françoise, Mille, Alain, & Pinon, Jean Marie. 1997.
Information retrieval on the world wide web using a decision making system.
Pages 284–295 of : Proceedings of the computer-assisted searching on the internet (riao 97).

References V

Csernel, B. 2008.
Résumé généraliste de flux de données.
Ph.D. thesis, ENST Paris.

Diday, E. 1971.
Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques.
Revue de statistique appliquée, **19**(2), 19–33.

Diday, E. 1975.
Classification automatique séquentielle pour grands tableaux.
Revue française d'automatique, informatique et recherche opérationnelle (rairo), 29–61.

References VI

Dubes, Richard C. 1987.
How many clusters are best ?—an experiment.
Pattern recogn., **20**(6), 645–663.

Elemento, O. 1999.
Apport de l'analyse en composantes principales pour l'initialisation et la validation de cartes topologiques de kohonen.
In : Actes des 7èmes journées de la société francophone de classification (sfc'99).

References VII

Ganti, Venkatesh, Gehrke, Johannes, Ramakrishnan, Raghu, & Loh, Wei-yin. 1999.
A framework for measuring changes in data characteristics.
Pages 126–137 of : In pods.
ACM Press.

Ganti, Venkatesh, Gehrke, Johannes, & Ramakrishnan, Raghu. 2000.
Demon : Mining and monitoring evolving data.
Pages 439–448 of : Ieee transactions on knowledge and data engineering.

References VIII


Halkidi, Maria, & Vazirgiannis, Michalis. 2001.
Clustering validity assessment : Finding the optimal partitioning of a data set.
Pages 187–194 of : Icdm '01 : Proceedings of the 2001 ieee international conference on data mining.
Washington, DC, USA : IEEE Computer Society.


Hardy, A. 2009.
Vadidation of clustering structure for symbolic data.
Cladag 2009.

References IX

Jaczynski, Michel, & Trousse, Brigitte. 1999.
Broadway : A case-based system for cooperative information browsing on the world-wide-web.
Pages 264–283 of : Collaboration between human and artificial societies, coordination and agent-based distributed computing.
London, UK : Springer-Verlag.

Kohonen, Teuvo. 1995.
Self-organizing maps. Third edn.
Springer Series in Information Sciences, vol. 30.
Springer.
Last edition published in 2001.

References X


Lebart, L., Morineau, A., & Piron, M. 1995.
Statistique exploratoire multidimensionnelle.
Dunod.


Liu, Bing, Ma, Yiming, & Lee, Ronnie. 2001.
Analyzing the interestingness of association rules from the temporal dimension.
Icdm '01 : Proceedings of the 2001 ieee international conference on data mining, 377–384.


Malek, Maria, & Kanawati, Rushed. 2001.
Cobra : A cbr-based approach for predicting users actions in a web site.
Case-based reasoning research and development : 4th international conference on case-based reasoning (iccbr 2001), 336–346.

References XI

Masseglia, Florent, Tanasa, Doru, & Trousse, Brigitte. 2004.
Diviser pour découvrir. une méthode d'analyse du comportement de tous les utilisateurs d'un site web.
Ingénierie des systèmes d'information, **9**(1), 61–83.

Milligan, G. W., & Cooper, M. C. 1985.
An examination of procedures for determining the number of clusters in a data set.
Psychometrika.

Murtagh, F. 1995.
Interpreting the kohonen self-organization feature map using contiguity-constrained clustering.
Pattern recognition letters, **16**, 399–408.

References XII


Neill, Daniel B., Moore, Andrew W., Sabhnani, Maheshkumar, & Daniel, Kenny. 2005.
Detection of emerging space-time clusters.
Pages 218–227 of : Grossman, Robert, Bayardo, Roberto J., & Bennett, Kristin P. (eds), *Proceedings of the eleventh acm sigkdd international conference on knowledge discovery and data mining, chicago, illinois, usa, august 21-24, 2005*.
ACM.


Noirhomme-Fraiture, Monique. 2000.
Multimedia support for complex multidimensional data mining.

Workshop on multimedia data mining, kdd'2000, 54–59.

References XIII

Spiliopoulou, Myra, Ntoutsi, Irene, Theodoridis, Yannis, & Schult, Rene. 2006.
Monic : modeling and monitoring cluster transitions.
Pages 706–711 of : Eliassi-Rad, Tina, Ungar, Lyle H., Craven, Mark, & Gunopulos, Dimitrios (eds),
Proceedings of the twelfth acm sigkdd international conference on knowledge discovery and data mining, philadelphia, pa, usa, august 20-23, 2006.
ACM.

Williams, W. T., & Lance, G. N. 1965.
Logic of computer based intrinsic classifications.
Nature, **207**, 159–161.

References XIV

Yan, T.W., Jacobsen, M., Garcia-Molina, H., & Dayal, U. 1996.
From user access patterns to dynamic hypertext linkin.
Pages 1007–1014 of : Computer network and isdn systems, (proceedings of www 2005), vol. 28.

Youness, Genane, & Saporta, Gilbert. 2004.
Une méthodologie pour la comparaison de partitions.
Revue de statistique appliquée, **52**(1), 97–120.

Zehraoui, Farida, Kanawati, Rushed, & Salotti, Sylvie. 2004.
Casep2 : Hybrid case-based reasoning system for sequence processing.
Advances in case-based reasoning, **3155**, 449–463.

References XV

Zhang, Tian, Ramakrishnan, Raghu, & Livny, Miron. 1996.
Birch : An efficient data clustering method for very large databases.
Pages 103–114 of : Jagadish, H. V., & Mumick, Inderpal Singh (eds), *Proceedings of the 1996 acm sigmod international conference on management of data, montreal, quebec, canada, june 4-6, 1996*.
ACM Press.