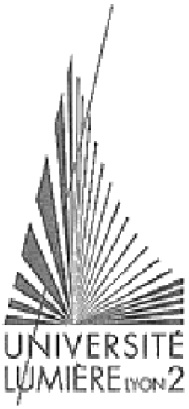
Regrouper les données textuelles et nommer les groupes à l'aide de classes recouvrantes

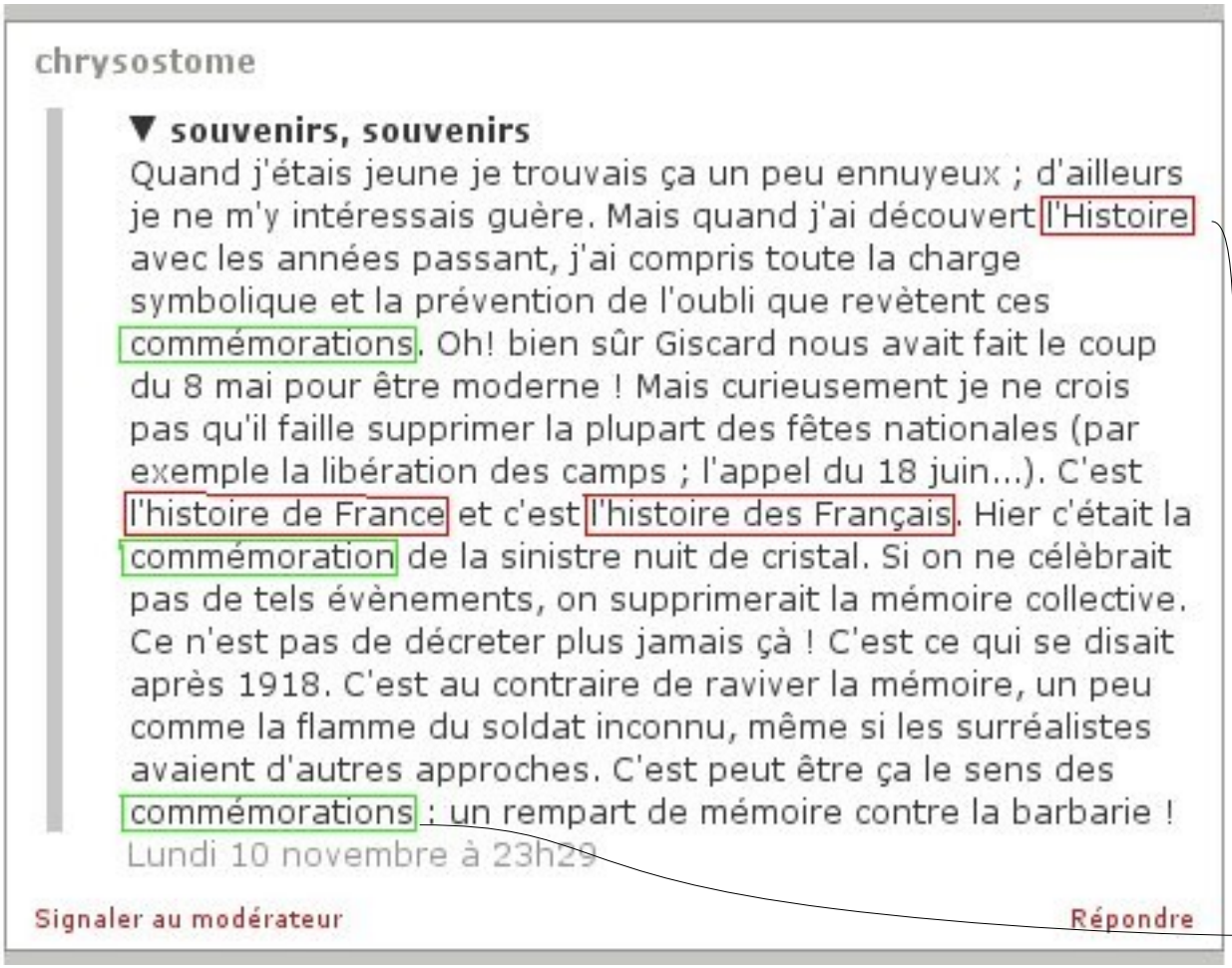
M-A. Rizoiu, J. Velcin et J-H. Chauchat

Laboratoire ERIC, Université Lumière Lyon 2, France



29 Janvier 2010 **Conférence EGC 2010**

Le problème


chrysostome

▼ **souvenirs, souvenirs**

Quand j'étais jeune je trouvais ça un peu ennuyeux ; d'ailleurs je ne m'y intéressais guère. Mais quand j'ai découvert l'Histoire avec les années passant, j'ai compris toute la charge symbolique et la prévention de l'oubli que revètent ces commémorations. Oh! bien sûr Giscard nous avait fait le coup du 8 mai pour être moderne ! Mais curieusement je ne crois pas qu'il faille supprimer la plupart des fêtes nationales (par exemple la libération des camps ; l'appel du 18 juin...). C'est l'histoire de France et c'est l'histoire des Français. Hier c'était la commémoration de la sinistre nuit de cristal. Si on ne célèbrait pas de tels évènements, on supprimerait la mémoire collective. Ce n'est pas de décreter plus jamais çà ! C'est ce qui se disait après 1918. C'est au contraire de raviver la mémoire, un peu comme la flamme du soldat inconnu, même si les surréalistes avaient d'autres approches. C'est peut être ça le sens des commémorations : un rempart de mémoire contre la barbarie !

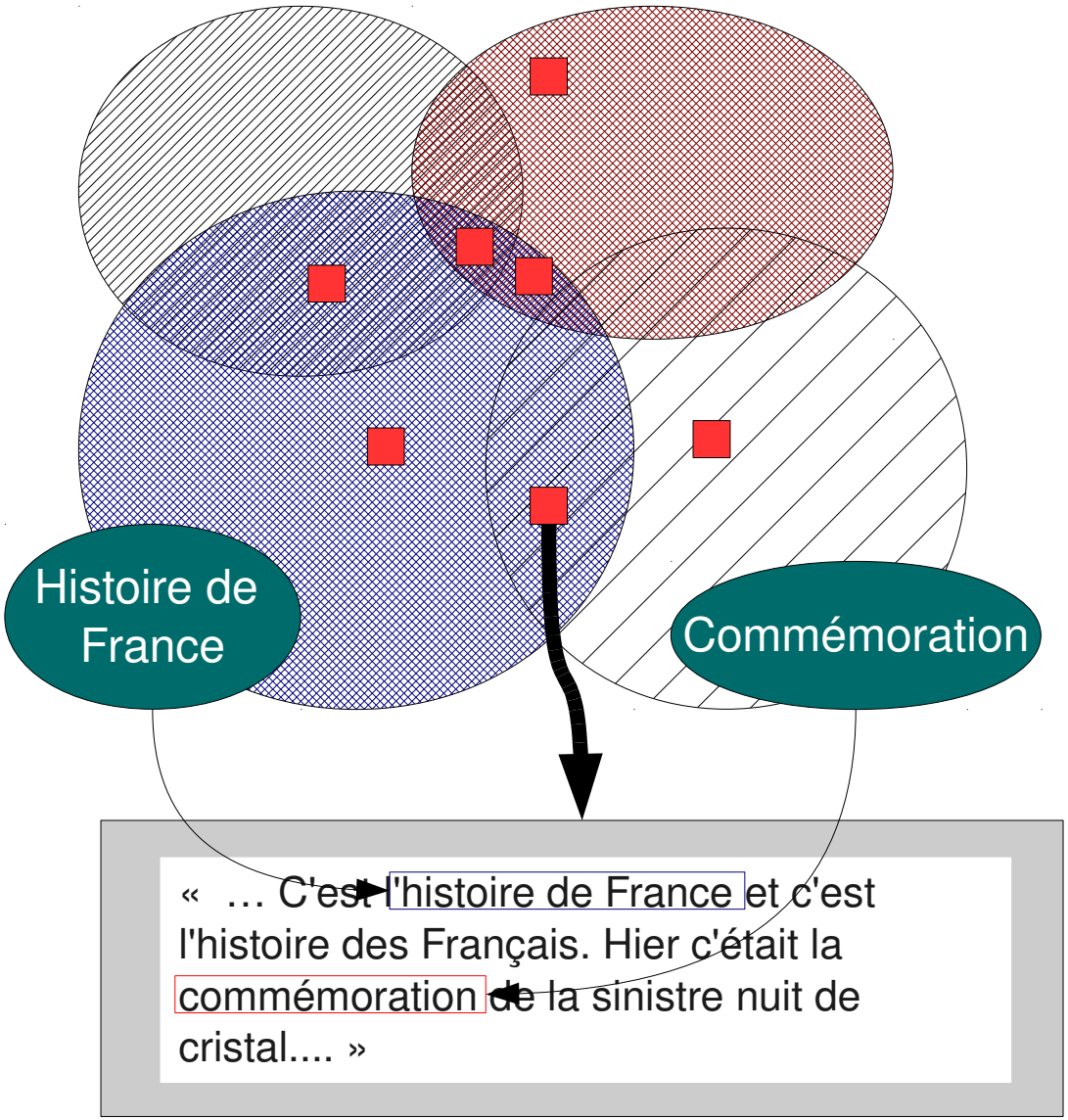
Lundi 10 novembre à 23h29

Signaler au modérateur Répondre

- Abondance d'information non-structurée
- Besoin de regrouper les documents par sujet
- **Textes en Langage Naturel**
- naturellement multi-thématiques

Un document,
plusieurs thématiques

Une approche



Apprentissage non-supervisé

Un algorithme automatique qui :

- regroupe avec recouvrement → **Overlapping Clustering**
- associe un nom qui synthétise bien le contenu de chaque groupe

Plan de la présentation

- Cadre du travail – état de l'art
- Notre approche
- Nos contributions
- Expérimentations
- Conclusions et perspectives

Overlapping Clustering

Singular Value Decomposition [Osinski 03]

- Décomposer la matrice terme / document en un produit de 3 matrices => type d'analyse factorielle qui réduit la dimension du problème
- Exprimer chaque document comme un produit pondéré des éléments de la base

Latent Dirichlet Allocation [Blei 03]

- Modèle mathématique constructif - mots générés par les « thématiques »
- Les documents - distributions probabilistes sur les « thématiques »

Overlapping KMeans (OKM) [Cleuziou 07, 08, 09]

- Une extension des KMeans qui permet le chevauchement
- Algorithme itératif qui minimise une fonction objectif

Nommer les groupes

Extraction de motif pertinents – 3 approches [Roche 04]

Linguistique

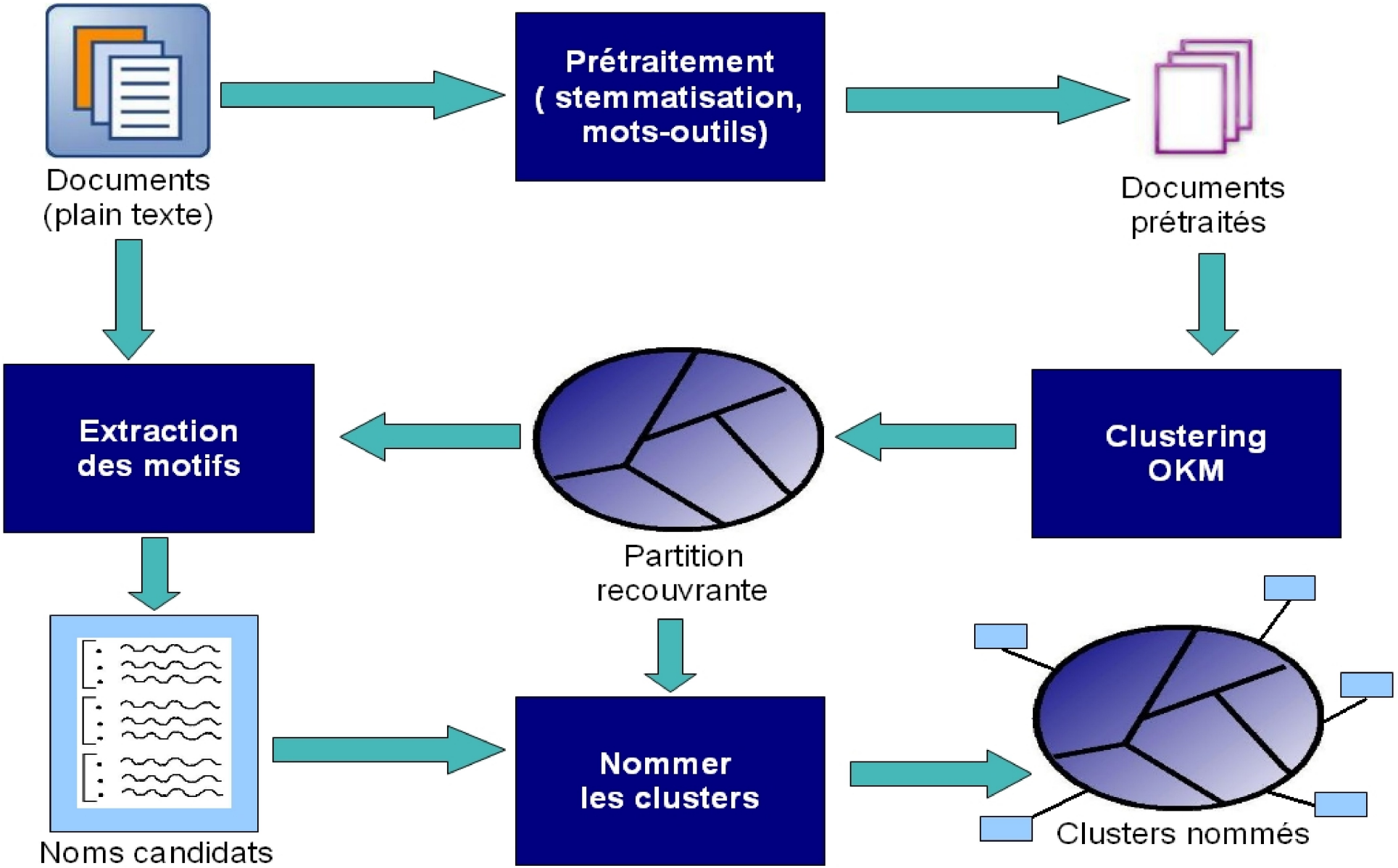
- Utilise des informations morphologiques et syntaxiques
- TERMINO [David et Plante 90], LEXTER [Bourigault 93];

Numérique / Statistique

- Utilise l'information statistique : Information Mutuelle, fréquence
- LOCALMAXS [Dias 00], EXIT [Roche 04], Suffix Tree [Zhang 01];

Hybride

- Ajoute des filtres linguistiques aux méthodes statistiques
- XTRACT [Smadja 91, Roche 04]


Documents
(plain texte)
Prétraitement
(stemmatisation,
mots-outils)
Documents
prétraités
Extraction
des motifs
Partition
recouvrante
Clustering
OKM
Noms candidats
Nommer
les clusters
Clusters nommés

Nos contributions

Une solution originale du problème d'extraction des thématiques en utilisant les recouvrement.

Une comparaison des performances du OKM et KMeans sur des textes en langage naturel en utilisant différents systèmes de pondération des termes (Fréquence de Termes, TFxIDF etc)

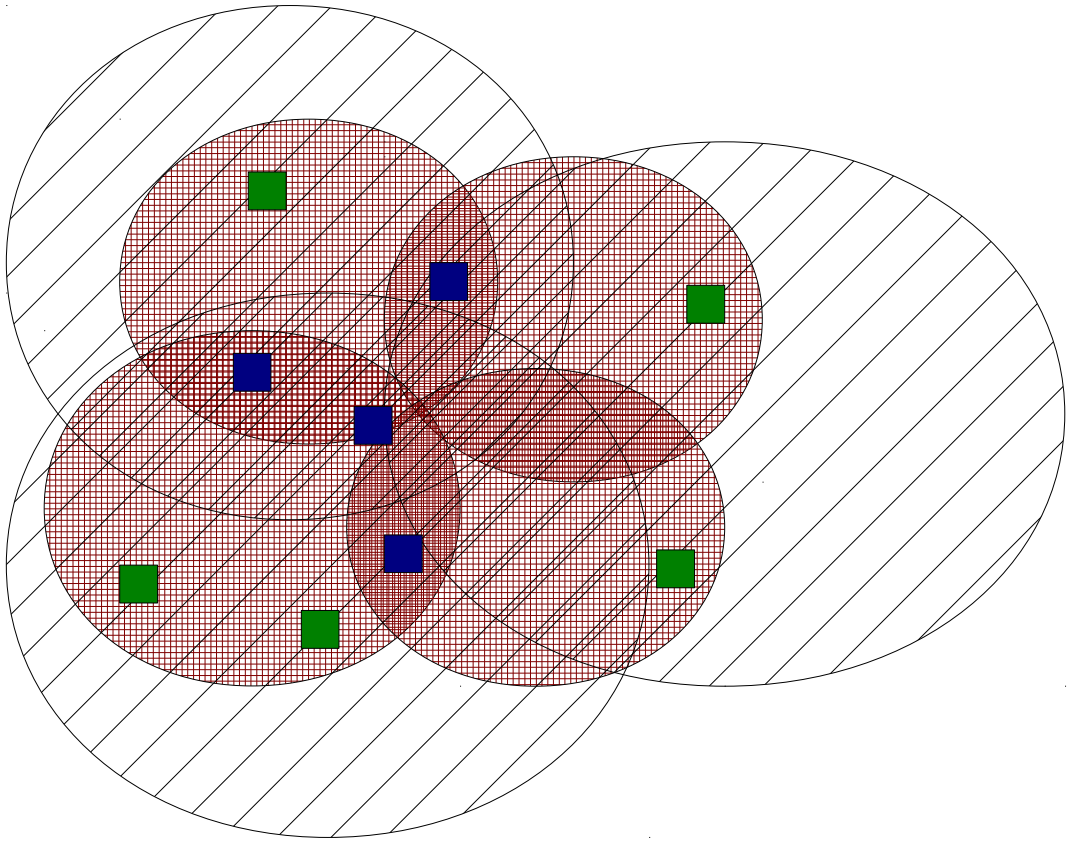
Une façon de résoudre partiellement le problème de l'initialisation au hasard et d'améliorer la stabilité de nos algorithmes. Une heuristique d'extraction des thématiques multi-niveaux.

Une véritable application qui sera utilisé dans le traitement des documents en ligne – liaison avec une entreprise : projet « Between »

Une conception modulaire qui permet :
- Le changement des composants
- L'étude du comportement des algorithmes de clustering et l'extraction des thématiques sur des données du monde réel

La difficulté de nommer les clusters recouvrants

- Un recouvrement important
- La plupart des noms candidats pertinents appartiennent a plusieurs clusters.
- Variance inter-cluster – mauvaix résultats



Prétraitement

Les buts:

- **Éliminer le bruit et la redondance**
- **Augmenter le pouvoir descriptif des mots**

« **Stemming** » - élimination des préfixes et suffixes, accents (Français) et lettres doubles

« **Élimination des stopwords** » - élimination des mot de liaison (préposition, article) sans grand pouvoir descriptif

OKM – Overlapping Kmeans [Cleuziou 07]

Clustering

Transformer les documents dans le **Modèle d'Espace Vectoriel**
– les mots sont les dimensions et les textes sont des vecteurs multidimensionnells.
– corrélation mot / texte → différents systèmes de pondération des termes (Présence/Absence, Term Frequency, TFxIDF)

Calculer la distance entre 2 vecteurs en utilisant la **distance du cosinus.**

Partition recouvrante – chaque document peut être associé à plusieurs groupes

Suffix Tree [Zhang 01]

Extraction des motifs pertinents

Motif (keyphrase) : un groupe de un ou plusieurs mots qui sont considérés pertinents s'ils sont ensemble (ex. « fouille de données »)

Motif pertinent : un motif qui est pertinent (il synthétise bien) le texte à partir duquel il a été extrait

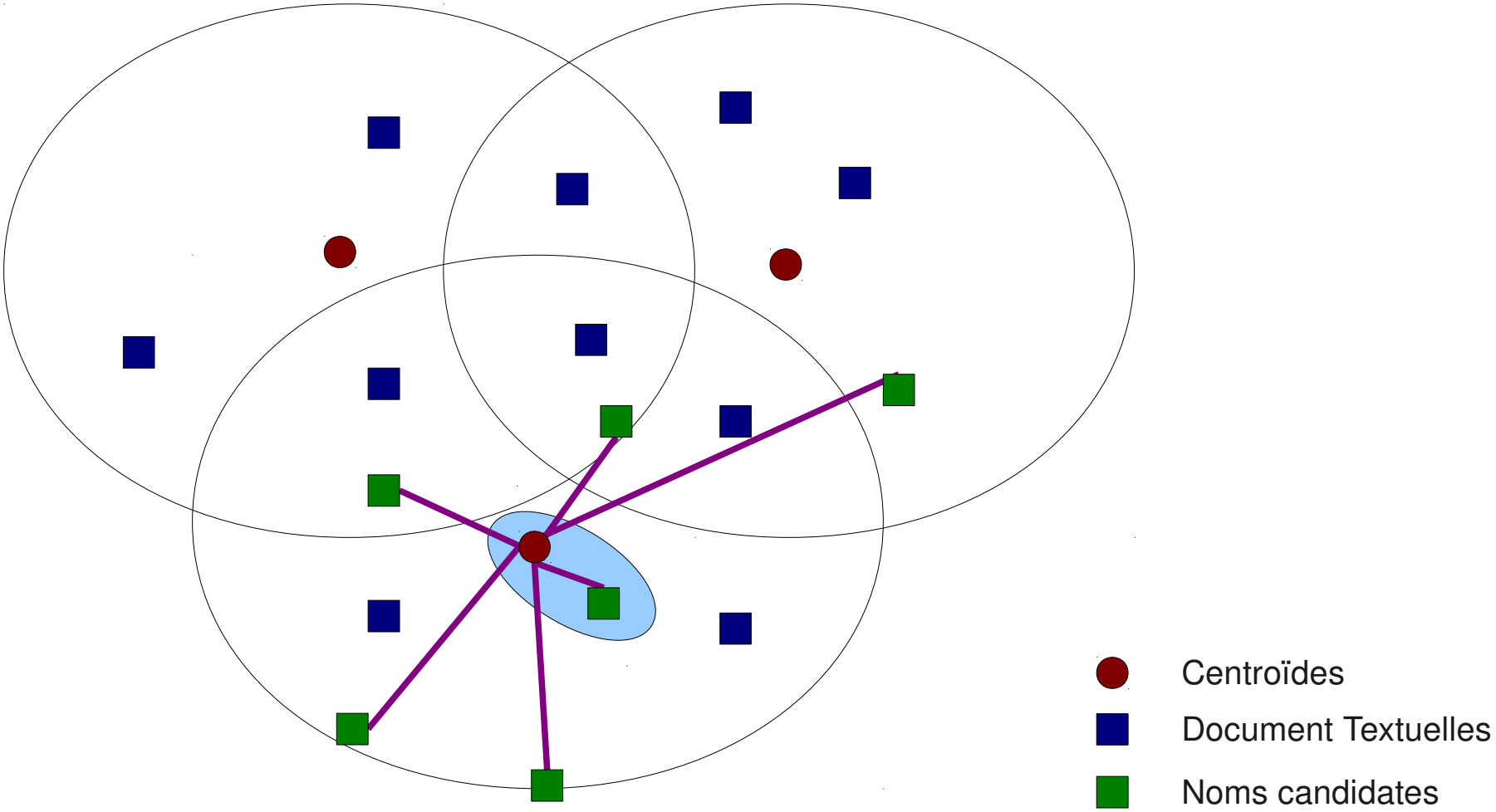
Algorithme basé sur la propriété de complétude des motifs.

Ex: « Président Nicolas » vs « Président Nicolas Sarkozy »

No	Suffix	Start Pos
1	a reunion	4
2	are having a reunion	2
3	having a reunion	3
4	reunion	5
5	we are having a reunion	1

L'arbre des suffixes pour la phrase « we are having a reunion »

Associer les noms aux groupes



Méthodologie

Associer les noms aux groupes

- Extraire des motifs pertinents à partir des clusters construits.
- Traduire le nom des candidats dans le Modèle d'Espace Vectoriel.
- Traiter les candidats comme des pseudo-documents et calculer la similitude avec les centroïdes.
- Choisi les motifs fréquentes plus proche des centroïdes.

-> Centroid[2]:
--> Top rated words: "comemo" "fete" "aniversai" "jambon" "bastil" "pris" "journ" "revolution" "juilet" "fed"
--> Name candidates: "commémoration" "jours de commémoration" "nombre de commémorations" "fête" "fête de la fédération" "fête nationale" "prise de la bastille" "anniversaire" "prise" "journée"
--> Cluster name: ***"commémoration"***

Exemple de sortie de notre algorithme: les mots les plus représentatif, les 10 premiers candidats et le nom choisi pour le cluster.

Comment ça marche?

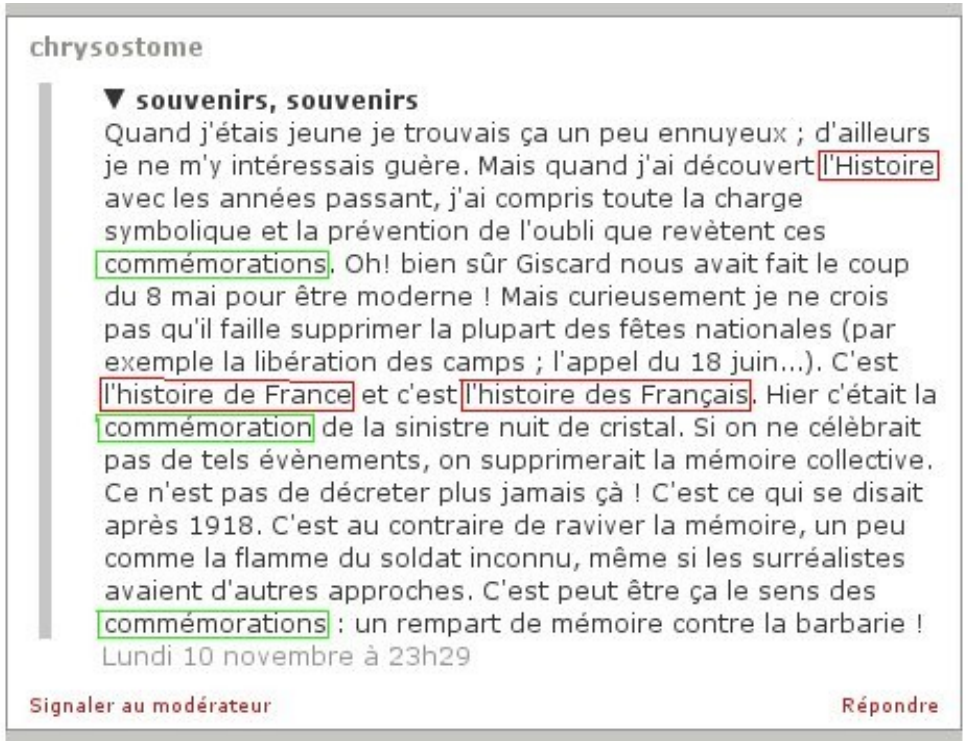
Un jeu de données réel : le forum « Y a-t-il trop de commémorations en France? », sur www.liberation.fr

Sortie du programme

--> Iteration no 11:
---> Objective function value: 189.154
---> Partitions:
----> Cluster 0 [101]: .....
----> Cluster 1 [90]: .....
*----> Cluster 2 [128]: ..... **texte_81** .....*
*----> Cluster 3 [192]: ..... **texte_81** ....*

Result - Cluster description:

-> Centroid[0]: "jours fériés"
-> Centroid[1]: "travailler plus pour gagner"
*-> Centroid[2]: **"commémoration"***
*-> Centroid[3]: **"histoire de france"***



chrysostome

▼ **souvenirs, souvenirs**

Quand j'étais jeune je trouvais ça un peu ennuyeux ; d'ailleurs je ne m'y intéressais guère. Mais quand j'ai découvert l'Histoire avec les années passant, j'ai compris toute la charge symbolique et la prévention de l'oubli que revêtent ces commémorations. Oh! bien sûr Giscard nous avait fait le coup du 8 mai pour être moderne ! Mais curieusement je ne crois pas qu'il faille supprimer la plupart des fêtes nationales (par exemple la libération des camps ; l'appel du 18 juin...). C'est l'histoire de France et c'est l'histoire des Français. Hier c'était la commémoration de la sinistre nuit de cristal. Si on ne célèbrait pas de tels évènements, on supprimerait la mémoire collective. Ce n'est pas de décreter plus jamais çà ! C'est ce qui se disait après 1918. C'est au contraire de raviver la mémoire, un peu comme la flamme du soldat inconnu, même si les surréalistes avaient d'autres approches. C'est peut être ça le sens des commémorations : un rempart de mémoire contre la barbarie !

Lundi 10 novembre à 23h29

Signaler au modérateur Répondre

Document « texte_81 » sur le site du forum

Demandes du monde réel

Problème de l'initialisation aléatoire

- Ajouter de l'information *a priori*;
- Injecter les motifs pertinents comme centroïdes initiaux et construire la partition autour d'eux.

jours fériés
travailler plus pour gagner
commémoration
histoire de france
jours de commémoration
bonne idée
américains

Extraction des thématiques multi-niveau

- Ajouter un deuxième niveau de thématiques pour affiner les résultats
- Réitérer l'algorithme sur les documents de chaque cluster – premier niveau

[2]: commémoration
|--> [0]: "leçon"
|--> [1]: "historique"
|--> [2]: "pouvoir"
|--> [3]: "vrai"

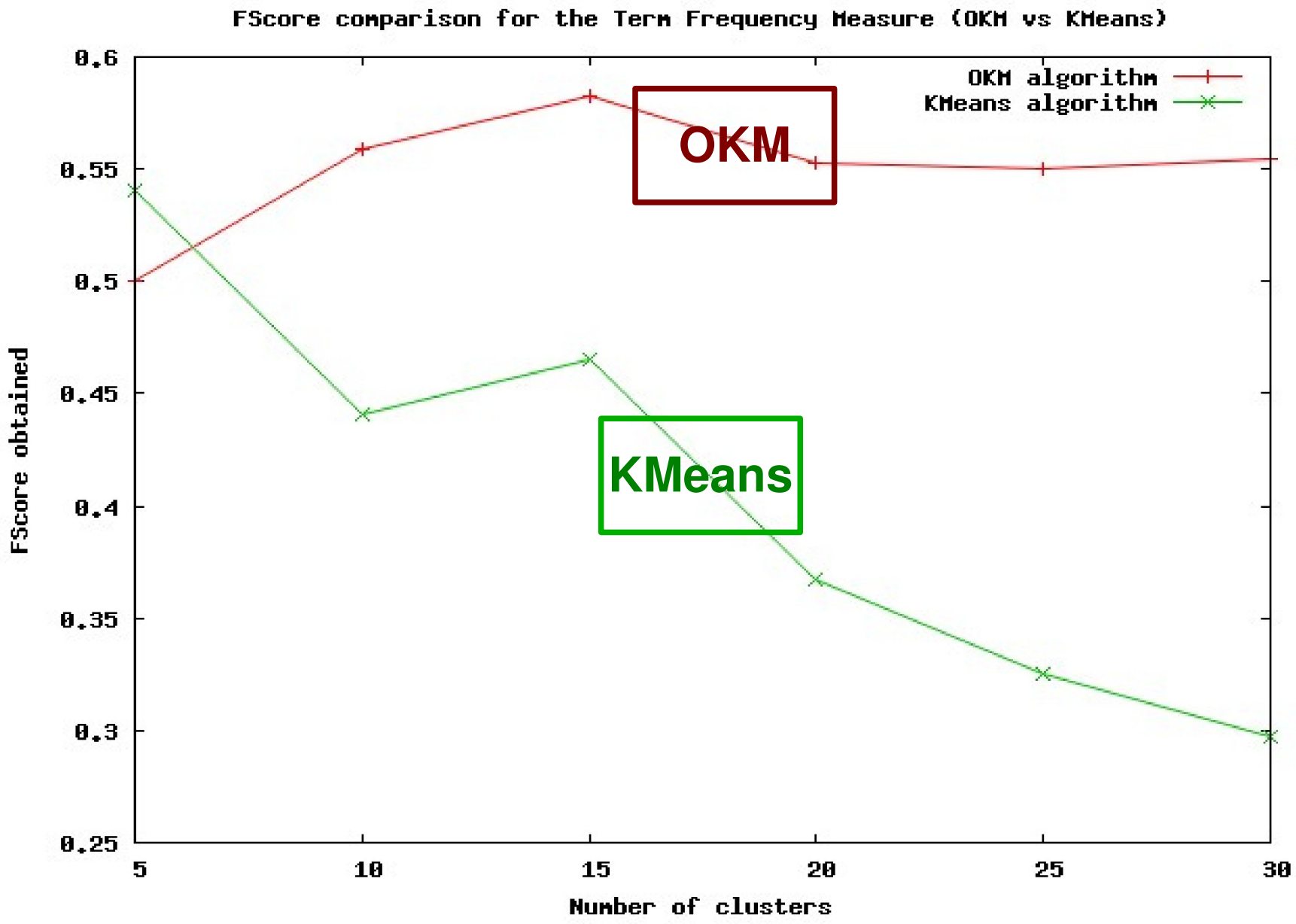
Expérimentations en 2 langues:

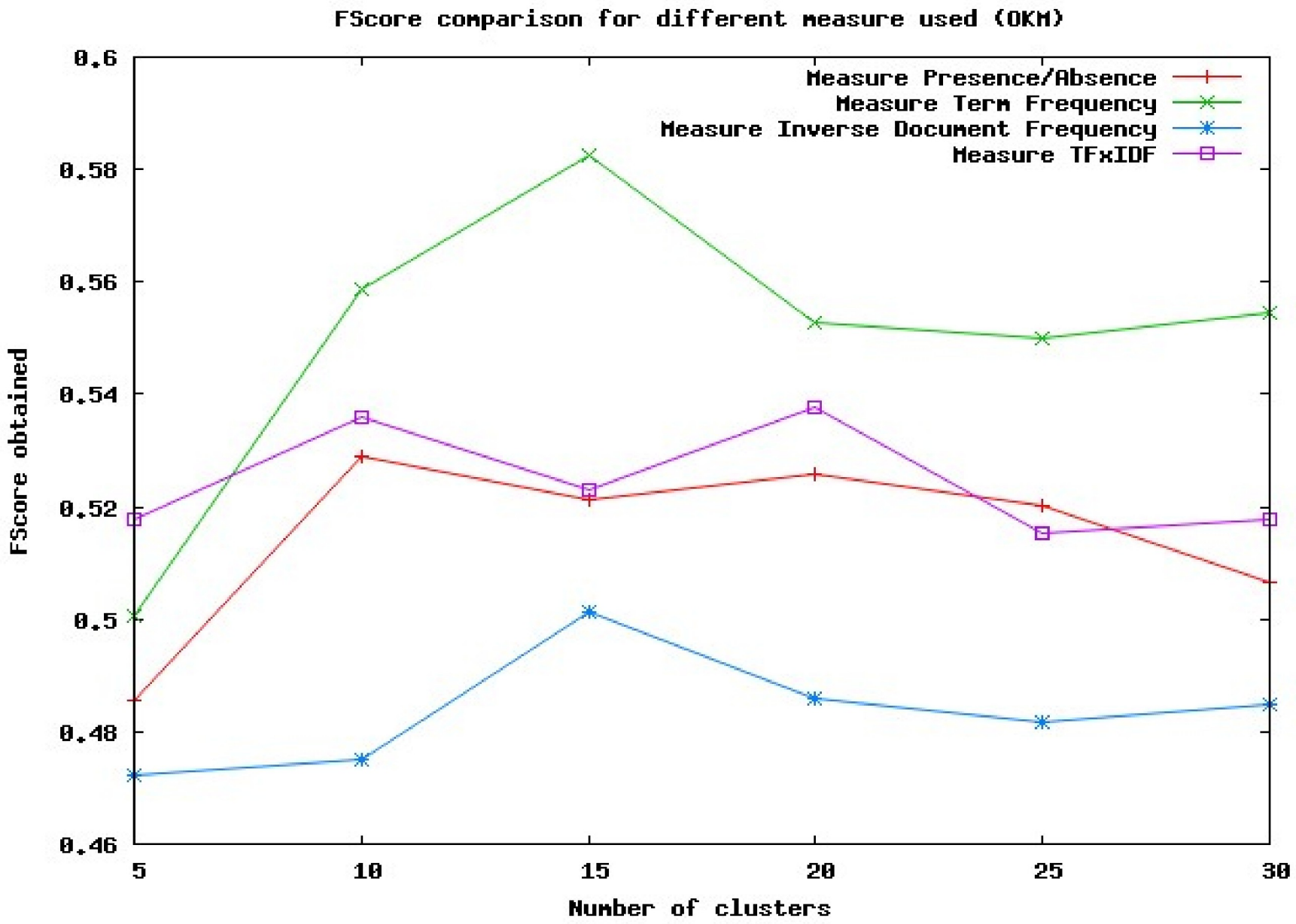
Français:

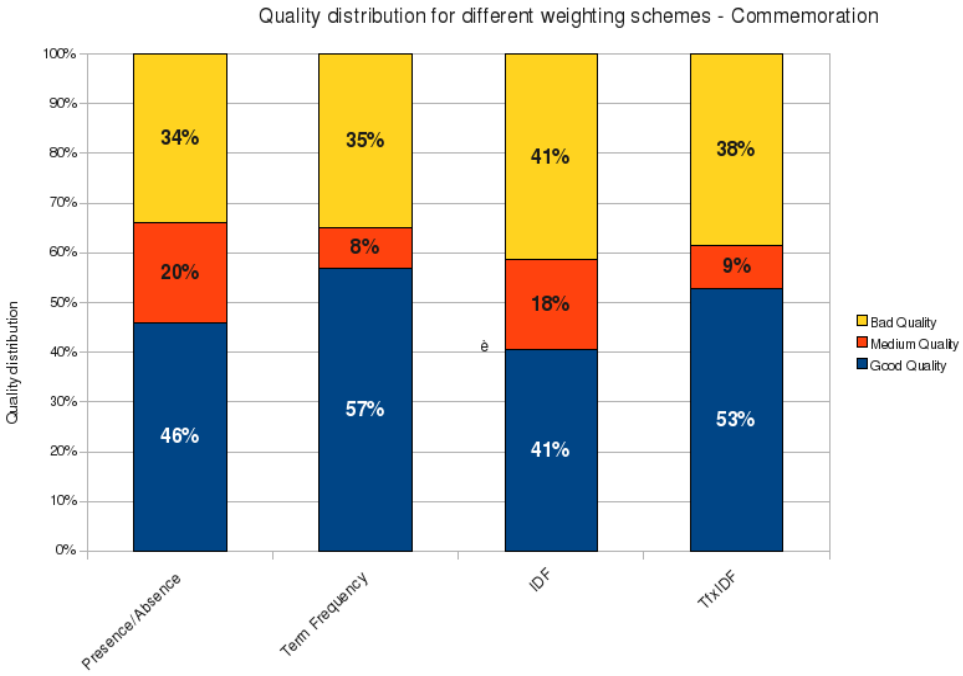
- Corpus du forum: « Y a-t-il trop de commémorations en France? », sur www.liberation.fr
- 276 documents courtes: 3 jusque 300 mots
- style d'écrire: informel
- utilisée dans le cadre de projet avec notre partenaire industriel

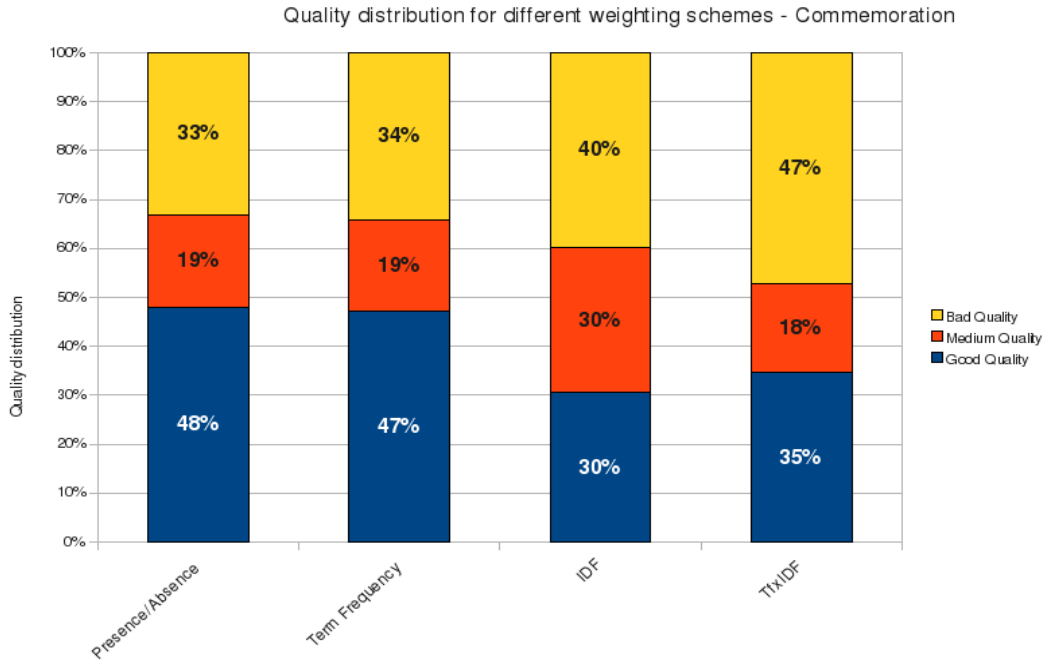
Anglais:

- Sous-partie du corpus « Reuters » - documents qui ont au moins une étiquette donnée par un expert
- 262 documents de dimension moyenne: 21 jusque 1000 mots
- style d'écrire: article de journal

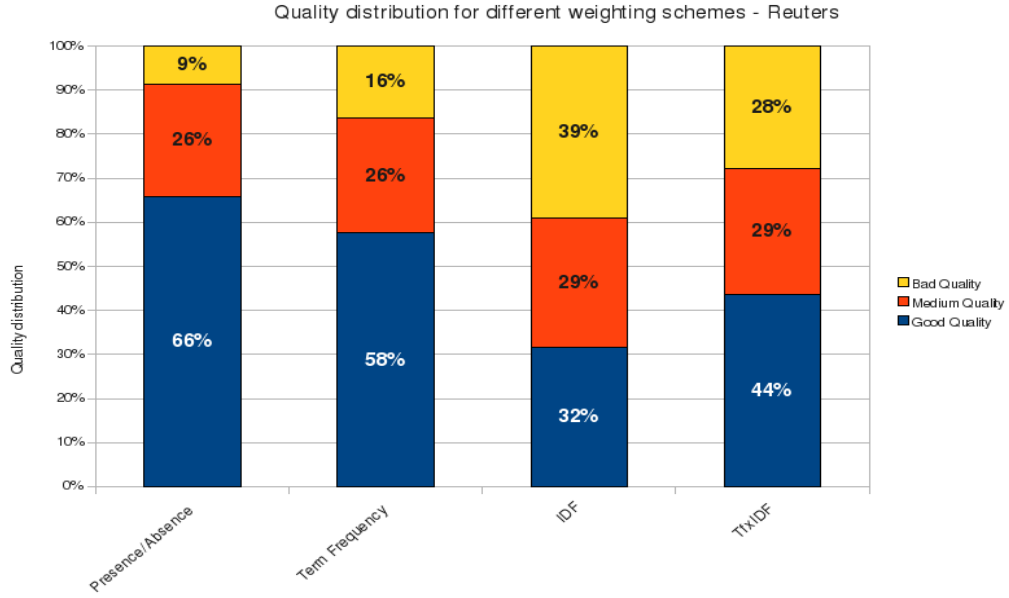

FScore comparison for the Term Frequency Measure (OKM vs KMeans)
OKM algorithm
KMeans algorithm
OKM
KMeans
FScore obtained
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
5
10
15
20
25
30
Number of clusters


FScore comparison for different measure used (OKM)
Measure Presence/Absence
Measure Term Frequency
Measure Inverse Document Frequency
Measure TFxIDF
FScore obtained
0.6
0.58
0.56
0.54
0.52
0.5
0.48
0.46
5
10
15
20
25
30
Number of clusters





En haut, résultats sur le corpus « **Commémoration** »



A droite, résultats sur le corpus « **Reuters** »

Conclusions

Un solution avec chevauchement - plus appropriée pour classer des données textuelles.

Étude des performances sur les données du monde réel

Une programmation modulaire qui permet d'introduire de nouveaux algorithmes.

Perspectives

Plus de comparaisons avec d'autres algorithmes de clustering « recouvrants » (Fuzzy, EM, etc)

Évaluation semi-automatisée des noms de cluster, qui va rendre possible d'évaluer automatiquement les noms extraits par le système.

Mis-a-jour de OKM vers wOKM [Cleuziou 09] → version pondérée

Améliorer la façon dont les noms sont associés à des groupes (WordNet, ontologies etc)

Bibliographie

[Osinski 03] Stanislaw Osinski. An algorithm for clustering of web search results. Master's thesis, Poznan University of Technology, Poland, 2003.

[Roche 04] M. Roche. Intégration de la construction de la terminologie de domaines spécialises dans un processus global de fouille de textes, 2004.

[Cleuziou 07] Guillaume Cleuziou. Okm : une extension des k-moyennes pour la recherche de classes recouvrantes, 2007

[Cleuziou 08] Guillaume Cleuziou and Jacques-Henri Sublemontier. Étude comparative de deux approches de classification recouvrante : Moc vs. okm, 2008.

[Cleuziou 09] Guillaume Cleuziou. Okmed et wokm : deux variantes de okm pour la classification recouvrante, 2009.

[David et Plante 90] S. David et P. Plante, « De la nécessite d'un approche morpho-syntactique dans l'analyse de textes », 1990

Bibliographie

[Bourigault 93] D. Bourigault « Analyse syntactique locale pour le repérage de termes complexes dans un texte », 1993

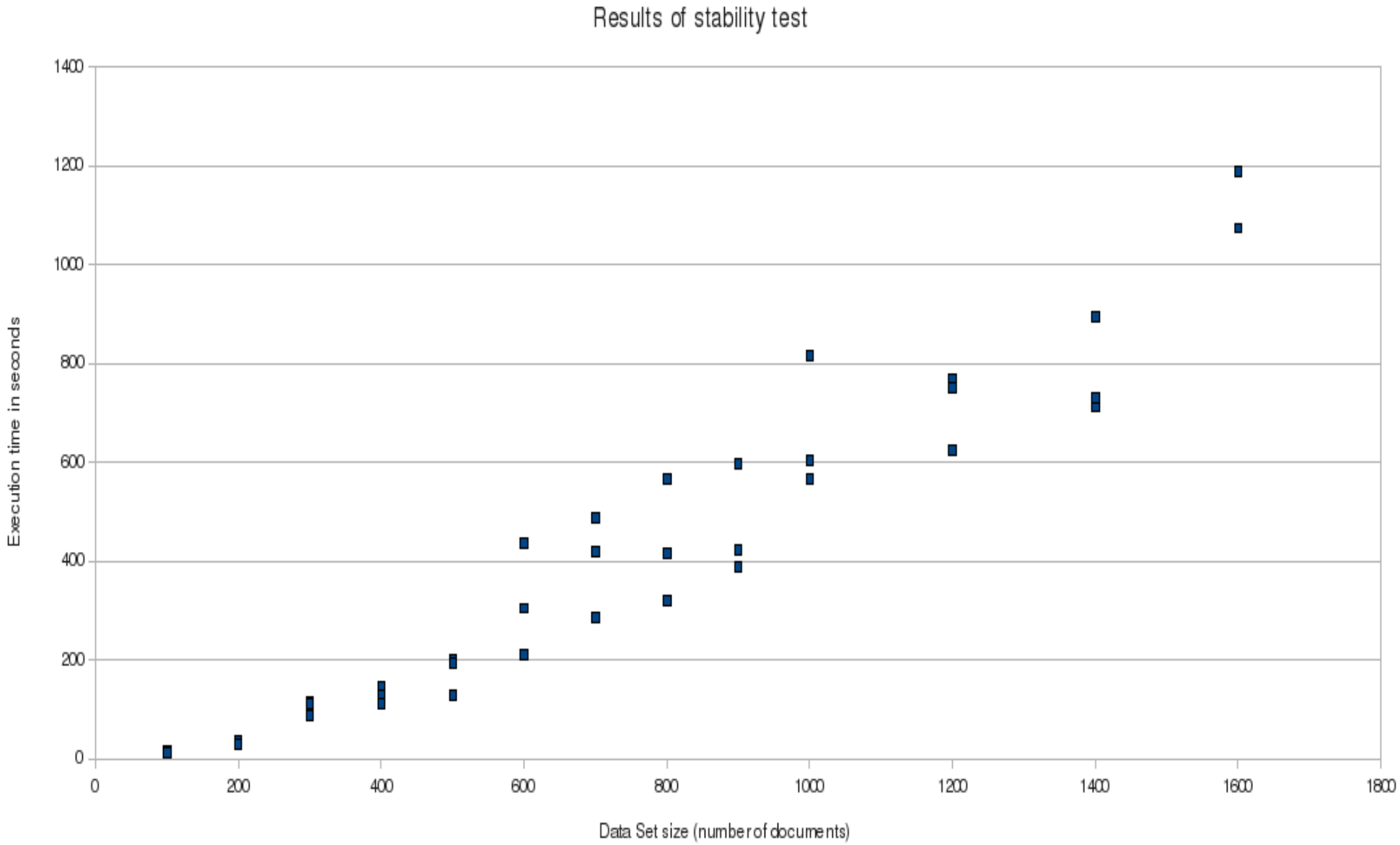
[Dias 00] G. Dias, S. Guillore et J.G. Pereira Lopes, « Extraction automatique d'associations textuelles a partir de corpora non traite », 2000

[Blei 03] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent dirichlet allocation. Journal of Machine Learning Research, 2003.

[Smadja 91] Frank A. Smadja. From n-grams to collocations: an evaluation of xtract, 1991

[Zhang 01] Dell Zhang and Yisheng Dong. Semantic, Hierarchical, Online Clustering of Web Search Results, 2001

Je vous remercie !!


Results of stability test
Execution time in seconds
Data Set size (number of documents)
0
200
400
600
800
1000
1200
1400
1600
1800