UQAC
Université du Québec
à Chicoutimi

Devoir #2 : Clustering et Schéma d'association pour la fouille de texte

Session : Automne2014

Forage de données
(8INF954)

Département d'informatique et de mathématiques

Présenté à : Professeur: A.Bouzouane

Travail de : AMAMOU Houssem

Date de remise : 28/10/2014

Synthèse de l'article :

Ce papier permet d'introduire une nouvelle méthode de classification de texte. En utilisant des algorithmes d'associations de règles, des ensembles utilisés fréquemment peuvent être découvert de façon plus efficace. Ces ensembles sont composés de termes. Cette approche a l'avantage de réduire considérablement la taille des données. Trois critères doivent être pris en compte dans la classification de texte : la grande dimensionnalité des données, la grande taille des bases de données, une description sémantique des clusters. Grâce à une étude réalisée, les chercheurs ont conclu que k-means est plus performant que les méthodes hiérarchiques et plus particulièrement, une variante de cet algorithme qui est bisecting k-means est encore plus performante. Les algorithmes utilisant la fréquence des items est une façon naturelle de procéder car il permet de réduire la dimensionnalité des documents. Dans les approches classiques, le « *standard overlap* (SO)» est utilisé, il permet de calculer le chevauchement entre deux clusters en utilisant la couverture de chaque cluster plus cette valeur de chevauchement est proportionnelle par rapport au taux d'erreur cependant cette approche favorise les ensembles ayant un nombre réduit d'items. L'alternative proposée dans ce papier est d'utiliser l'entropie grâce à l'inverse de la mesure de distribution d'un document et on définit le « *entropy overlap (EO)*» qui calcule la distribution des documents dans un cluster. Grâce aux tests, cette nouvelle approche donne une classification de meilleure qualité. Cette approche consiste à calculer tous les overlap pour tous les ensembles fréquents, extraire l'ensemble avec le minimum d'overlap, enlever les documents que couvre cet ensemble. FTC permet de générer une description naturelle à travers les clusters générés grâce aux termes fréquents. HFTC génère des clusters hiérarchiques qui sont plus compréhensibles que les résultats d'autres algorithmes.

Travail effectué :

L'algorithme FTC et HFTC a été implémenté dans une application externe à Weka. Grâce à son interface usager il est possible de paramétrer cette application. Comme demandé la valeur de minsup peut être régler manuellement. Elle calculer en % pour une meilleure compréhension. L'utilisateur a le choix entre les deux heuristiques proposées dans l'article : le recouvrement standard et l'entropie de recouvrement. L'application qui a été développé comporte 6 classes. La classe principale est la classe Ftc qui implémente les deux algorithmes FTC et HFTC. Une deuxième classe Data a été implémenté, elle permet de représenter les documents de la base des données. La classe overlap est capable de calculer l'overlap de chaque ensemble de termes fréquents selon le choix de l'heuristique sélectionné par l'utilisateur. La classe FrequentItem représente les ensembles de termes fréquents. Afin de tester la validité de l'application la base de données de Reuters a été utilisée, c'est un ensemble composé de 21578 documents mais comme indiqué dans l'article seul à peu près 8600 ont été sélectionnée pour les tests. Afin de lancer l'application il suffit juste d'indiquer le dossier dans lequel se trouvent les

documents à traiter. Nous allons présenter ci-dessous quelques copies d'écran de l'application réalisée et le résultat des tests sur la base de données Reuters.

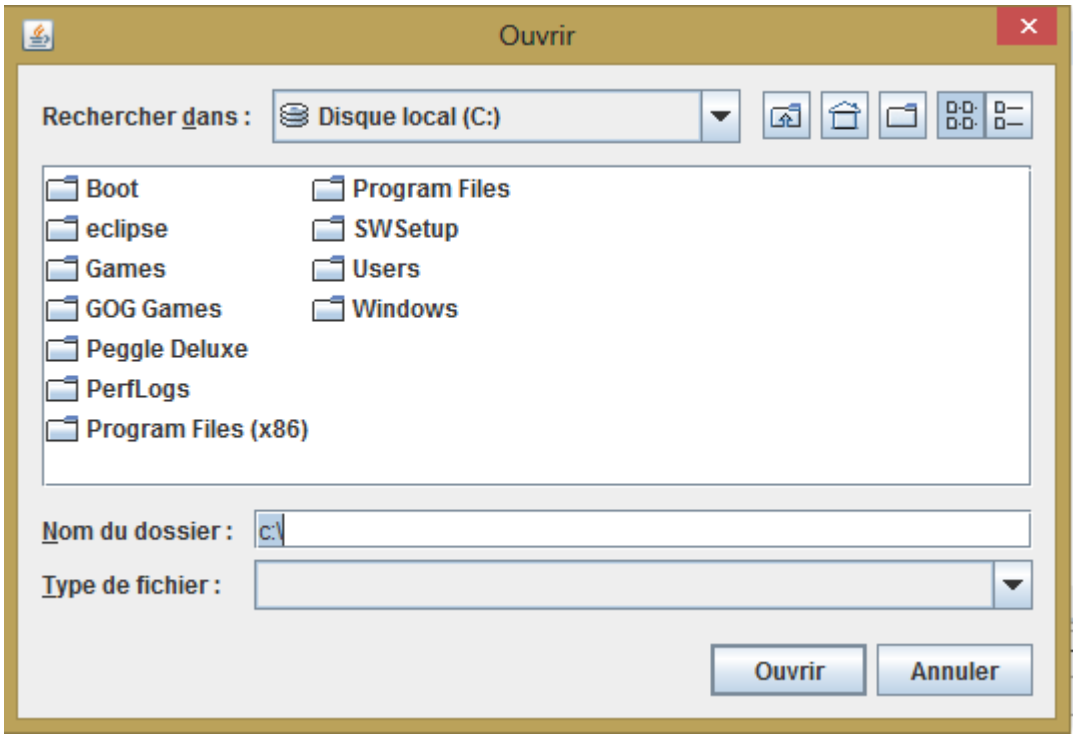

Figure 1 : Menu pour le choix du répertoire principale

Cette première capture d'écran (Fig. 1) met en évidence la possibilité pour un utilisateur de choisir le dossier dans lequel se trouvent les documents à traiter

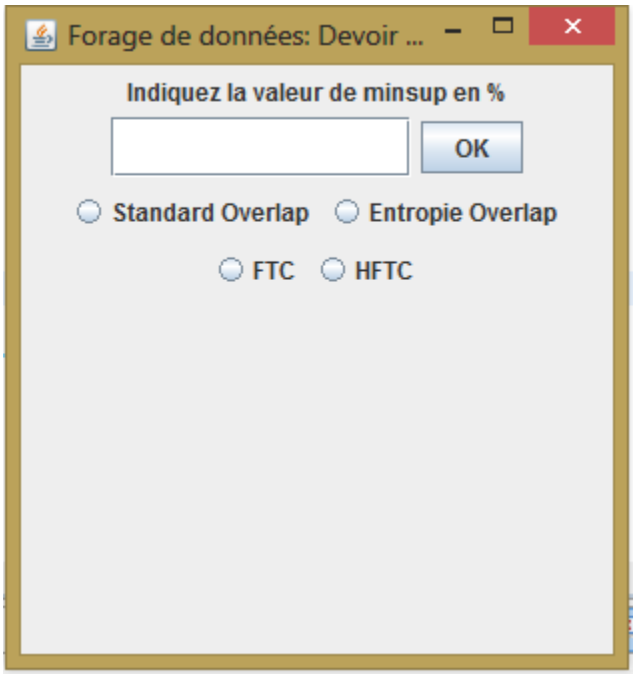

Figure 2 : Menu pour le choix des paramètres de l'application

La figure 2 est le menu principale de l'application, grâce à celui –ci l'utilisateur peut indiquer la valeur de minsup, de choisir l'heuristique avec laquelle il veut travailler c'est-à-dire soit l'overlap standard ou l'overlap entropique.

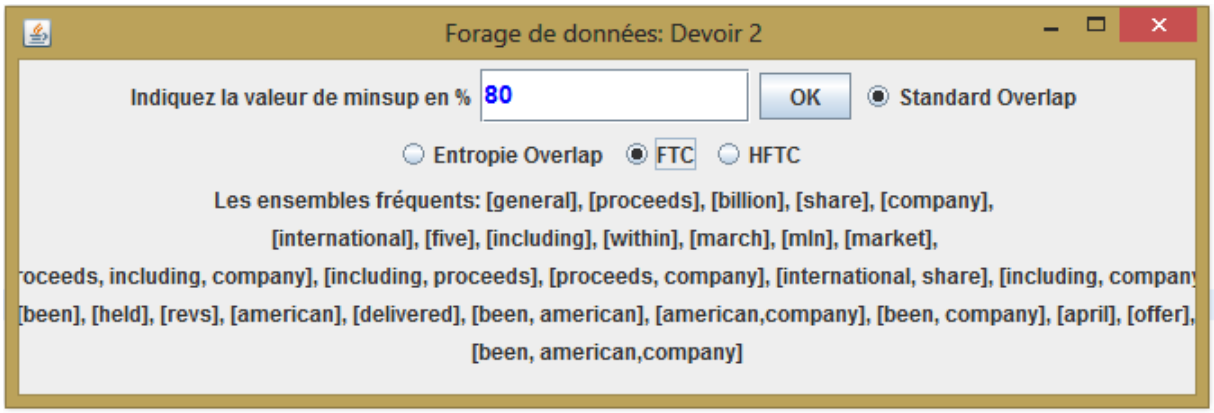

Figure 3 : Un test réalisé avec les paramètres indiqués

La figure 3 permet de montrer un test réalisé avec les paramètres tels que le choix du standard overlap comme heuristique et de FTC comme algorithme, la valeur de minsup est égale à 80%. Le résultat affiché est l'ensemble des ensembles des termes fréquents.

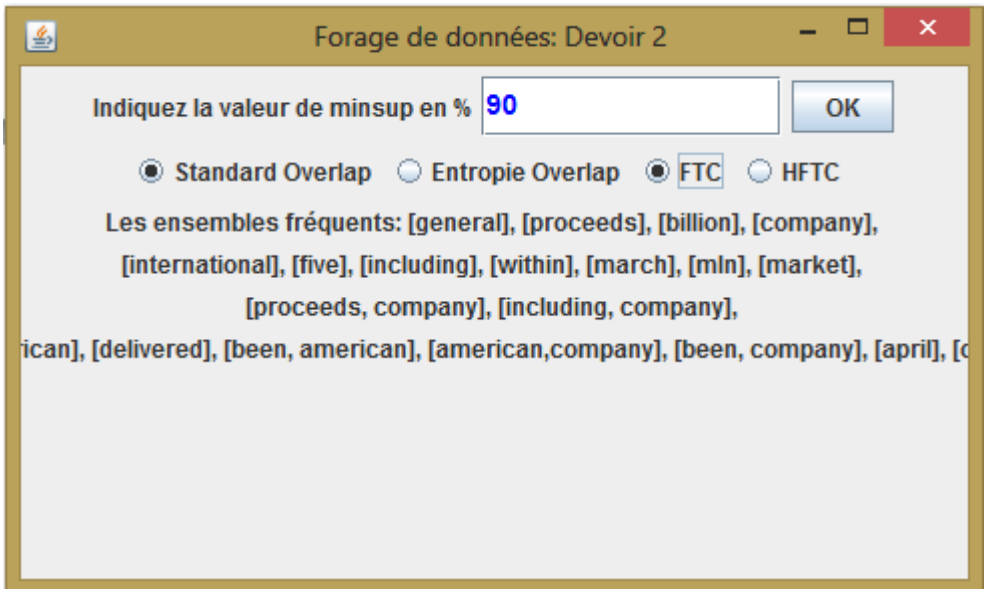

Figure 4 : Un test réalisé avec les paramètres indiqués

La figure 4 permet de mettre en valeur un test réalisé avec minsup égale à 90%.