EvalECD'2010

Évaluation des méthodes d'Extraction de Connaissances dans les Données

Organisateurs :

Fatiha Saïs (LRI, INRIA-Saclay, Université Paris-Sud – CNRS UMR8623)
Nicolas Béchet, Mathieu Roche (LIRMM, Université Montpellier 2 – CNRS UMR5506)

Vers une méthode d'évaluation de Système Interactif d'Aide à la Décision basé sur le processus d'ECD

Rania Bahloul*, Mounir Ben Ayed*
Adel M Alimi *

*REGIM: REsearch Group on Inteligent Machines Ecole nationale d'ingenieurs de sfax
Route Sokra km 3.5 BP W 3038 sfax. tunisie
{Rania.Bahloul, Mounir.benayed, Adel.Alimi}@ieee.org

Résumé. L'évolution des systèmes informatique utilisant de grande quantité de données mène à concevoir des systèmes interactifs faisant intervenir l'utilisateur. Ainsi, dans cet article nous proposons une méthode, centrée utilisateur, d'évaluation de Système Interactif d'Aide à la Décision (SIAD) basé sur le processus d'Extraction de Connaissances à partir de Données (ECD). Nous tenons en compte des deux métriques issues du domaine de l'Interface Homme Machine (IHM) « utilisabilité » et « utilité ». Cette méthode vise à intégrer l'évaluation de l'interface réalisée ainsi que les données obtenues à partir de chaque phase tout au long de ce processus pour aboutir aux données utiles permettant d'extraire une connaissance.

1 Introduction

L'évolution croissante des systèmes informatiques dans divers domaines et pour différents types d'usagers mène à concevoir des systèmes interactifs. Ces systèmes permettent l'intervention de l'utilisateur. Ils utilisent un outil décisionnel, pour aider à la prise de décision. Pour la prise de cette décision, il s'avère nécessaire d'extraire des informations utiles et utilisables par l'observation et l'analyse des données. Mais cela est très difficile à réaliser à cause de l'augmentation des quantités de données stockées dans des bases de données. Ainsi l'utilisation d'un outil décisionnel est indispensable à savoir le Data Mining. En puissant dans des bases de données volumineuses les méthodes de fouille de données ont pour but de présenter à l'utilisateur final une information fiable et interprétable. A fin d'obtenir un système interactif une combinaison entre l'interaction homme machine et un outil décisionnel a été réalisée. Le système obtenu est nommé « Système Interactif d'Aide à la Décision (SIAD) ». Pour avoir un système fiable qui répond au besoin de l'utilisateur il faut l'évaluer en termes de la qualité de l'interface réalisée et de l'utilité des données obtenues par le système. Pour que cette évaluation soit rigoureuse, elle doit être faite tout au long de la conception et la réalisation. On parle d'évaluation à priori et à posteriori [4]. Cette évaluation doit être faite à chaque phase du processus d'extraction de connaissance à partir de données.

2 Définitions et objectifs du Data Mining :

L'exploration de données, aussi connue sous les noms fouille de données, data mining (forage de données) ou encore Extraction de Connaissances à partir de Données (ECD en français, KDD en Anglais), a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données ou de base de données, par des méthodes automatiques ou semi-automatiques, et l'utilisation industrielle ou opérationnelle de ce savoir.

La fouille de données est une étape centrale du processus d'ECD [2]. Elle correspond à l'ensemble des méthodes et des techniques issues de spécialités différentes : statistiques, algorithmes génétiques, réseaux de neurones. Qui a partir de données permettent d'obtenir des connaissances exploitables.

Les objectifs de la fouille de données peuvent être regroupés en cinq fonctions : classification, estimation, prédiction, optimisation, segmentation et explication.

2.1 Fonctionnement du data mining

Le processus d'ECD peut être redéfini par une suite d'opérations de transformation et d'analyse des données qui consiste à:

Poser le problème qui présente la première phase où on définit les objectifs, le résultat ainsi que les moyens de mesurer le succès de l'étape de datamining.

La recherche des données par association de multiple source de données sous une forme unique. Elles seront par la suite sélectionnées pour avoir des données ayant un rapport avec l'analyse demandée.

Ces données subiront une phase de nettoyage puis de transformation pour les préparer à la fouille de données.

Dans la phase de fouille de données, des méthodes intelligentes, pour extraire l'information pertinente, sont appliquées. Ces informations sont interprétées et évaluées pour extraire une connaissance.

3 Méthodes d'évaluation

L'évaluation est effectuée pour vérifier la qualité de l'interface du système à évaluer en se basant sur les critères ergonomiques et ceux imposées par l'utilisateur final de l'interface et l'utilité des résultats obtenus par le processus d'ECD.

Pour ce faire, nous avons opté pour l'évaluation des deux principaux critères, l'utilisabilité et l utilité.

3.1 Evaluation de l'utilisabilité

L'évaluation de l'utilisabilité est basée sur trois approches : approche analytique, approche centrée utilisateur et approche basée sur l'expert.

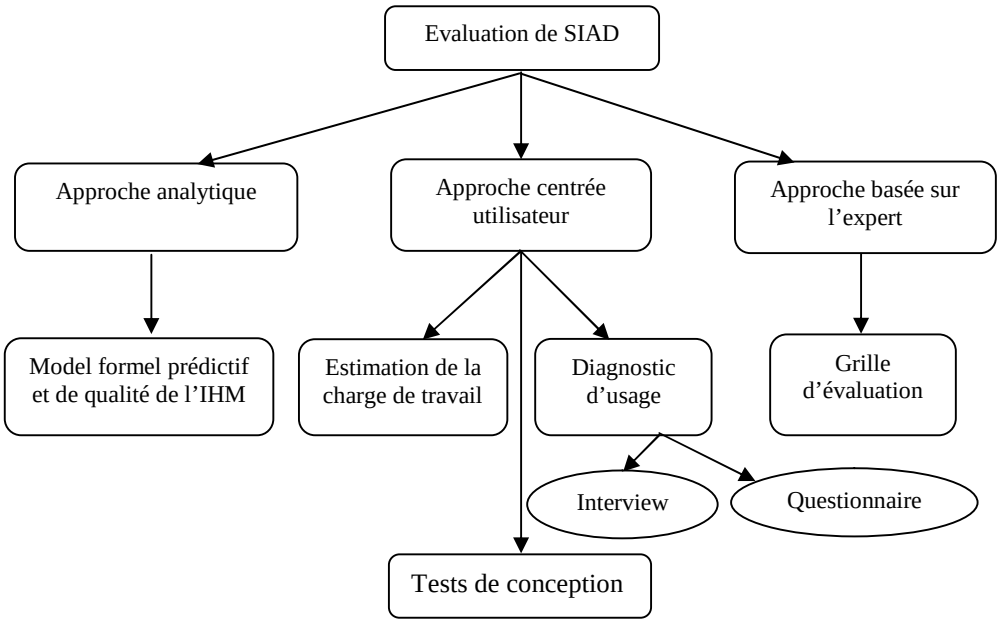


FIG. 1 – *Principe de la méthode d'évaluation*

3.1.1 Approche analytique

Il s'agit d'un diagnostic de détection des erreurs de conception en vue de fournir des alternatives de conception, ou d'évaluation permettant de déterminer jusqu'à quel point le système interactif est adapté pour les tâches pour lesquelles il a été conçu, ou encore d'une évaluation de conformité à des normes ?

Pour appliquer cette approche il faut utiliser un model de référence sur lequel se base la comparaison et par suite la validation de la qualité de l'IHM [9], dans ce travail nous avons fait recours aux qualités logicielle qui sont en rapport avec l'utilisateur tel que la validité, l'efficacité, et facilité d'emploi.

En plus de l'évaluation de l'utilisabilité et de l'utilité, nous avons évalué les critères ergonomiques en comparant l'interface du SIAD à évaluer avec les recommandations décrites par Scapin [9].

Alors que pour la prédiction, le concepteur, lors de la conception de l'IHM doit tenir compte du type des utilisateurs afin de générer des interfaces faciles à exploiter par tout type d'utilisateurs.

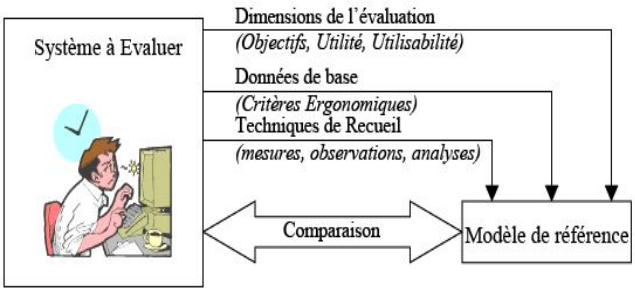


FIG. 2 – *Principe global de l'évaluation [10]*

3.1.2 Approche centrée utilisateur

Ces approches sont basées sur des techniques d'observation de l'utilisateur réel (utilisateurs finaux) et de recueil des données de l'interaction (questionnaire, interview, verbalisation, oculométrie, estimation de la charge de travail, etc.) afin d'analyser les traces de l'activité des utilisateurs. Ces approches permettent de détecter les problèmes réels que rencontre l'utilisateur lorsqu'il réalise sa tâche avec le système.

1. Estimation de la charge de travail

L'estimation de la charge de travail peut être utilisée à priori, lorsque l'interface est au stade du prototypage et à posteriori quand l'interface est réalisée. Cette estimation permet de mesurer le niveau de difficulté liée à l'utilisation de l'IHM.

Une telle estimation définie le rapport entre le temps nécessaire à l'utilisateur pour effectuer une tâche et le temps effectif pour l'exécuter.

2. Tests de conception

Les méthodes de tests de conception permettent d'évaluer et de valider un système interactif ou une IHM selon un cycle itératif tout au long du cycle de développement avec des utilisateurs pour cela pour l'évaluation du système durant son développement on a fait recours au processus unifié.

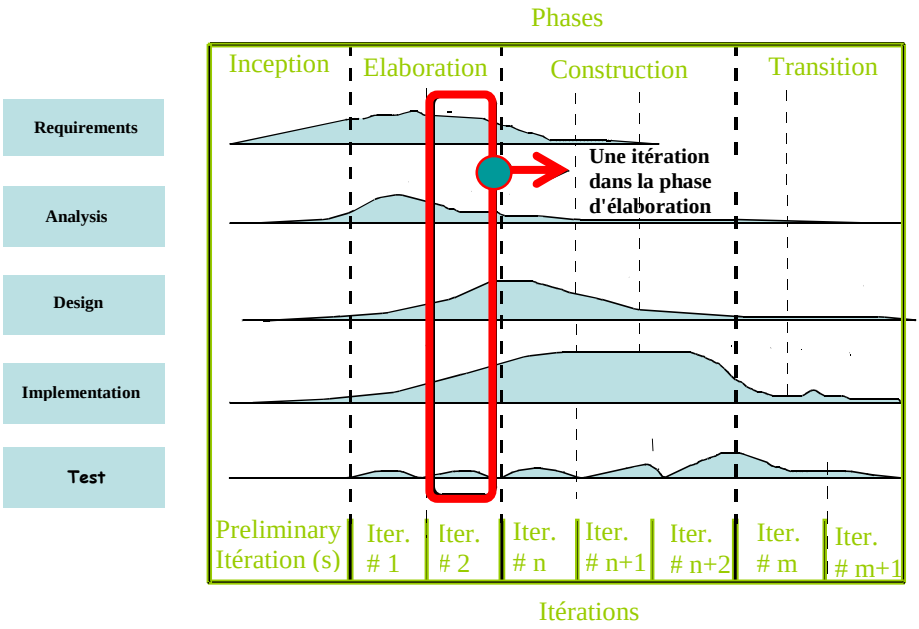


FIG. 3 – *Processus unifié* [3][5]

Ce processus comprend 4 phases :
- l'Inception : pour l'analyse de besoins, l'architecture générale du système et l'étude de faisabilité
- L'élaboration : permet de préciser les cas d'utilisation, de concevoir l'architecture du système et déterminer l'architecture de référence
- La construction : au cours de laquelle le logiciel est construit au moyen de plusieurs itérations et de nombreuses versions du système
- La transition : pour remettre le système aux utilisateurs finaux avec la mise en service et pour les former et les soutenir à l'utilisation. [3] [5]

Dans le cadre de notre travail, nous nous intéressons à l'évaluation des étapes des trois dernières phases car la phase d'inception ne comporte pas de test en effet,

- L'évaluation de la phase d'élaboration consiste à vérifier si le système à concevoir ainsi que son architecture rendent services à l'utilisateur
- Quant à la phase de construction, l'évaluation consiste à vérifier si le modèle de conception de l'IHM répond aux besoins de l'utilisateur
- Alors que pour la phase de transition l'évaluation consiste à tester le système réalisé pour identifier les anomalies

Nous avons utilisé ce processus pour la méthode de tests de conception car il permet de générer des modèles d'interface (ou prototype) lors de l'étape d'analyse qui seront évaluer en intégrant l'utilisateur pour exprimer son aptitude en vers l'interface du système en terme de

facilité d'apprentissage, de cohérence, d'efficacité et de prévention à l'erreur pour accomplir la tâche à réaliser par le système.

3. diagnostic d'usage

Cette approche n'est possible que si l'IHM est opérationnelle, et est prête pour être présentée aux utilisateurs finaux (des exceptions pour les interfaces qui sont dans une phase bien avancée de leurs développements peuvent être faites). L'évaluation se base essentiellement sur un recueil d'informations à travers l'interview ou le questionnaire.

Le questionnaire forme un document structuré ayant pour objectif de recueillir un ensemble d'appréciation et d'opinion sur l'attitude de l'utilisateur. Il peut aborder les aspects liés au fonctionnement du système et à l'ergonomie de l'interface.

3.1.3 Approche basée sur l'expert

Pour cette approche, l'évaluation est effectuée par un expert en ergonomie ou par un spécialiste en communication homme machine. Elle consiste à comparer les attributs et caractéristiques de l'interface à des recommandations ergonomiques ou normes pour détecter les erreurs de conception. Elles peuvent être combinées avec les approches centrées utilisateurs et permettent de juger la qualité ergonomique de l'interface.

1. Méthode d'inspection de l'utilisabilité

Cette méthode vise la détection des aspects des interfaces pouvant entraîner des difficultés d'utilisation qui consiste en la conformité à des recommandations ergonomiques qui consiste à juger la conformité des éléments de l'interface par rapport à des règles disponibles dans différents documents prenant généralement la forme de recueils. Pour ce faire nous avons comparé l'interface du système réalisé dans un milieu hospitalier avec les recommandations citées par L.Scapin et J.M Christian [8]

2. grilles d'évaluation

Les grilles d'évaluation permettent d'évaluer l'interface selon plusieurs critères ergonomiques.

Pour élaborer la grille d'évaluation qui nous permet d'aboutir à l'évaluation du système, nous avons utilisé le questionnaire réalisé dans l'approche centrée utilisateur de diagnostic d'usage. Ce questionnaire est exploité pour l'évaluation des quatre critères principaux Facilité d'apprentissage, Cohérence, Efficacité d'utilisation et Prévention et correction d'erreurs dans lesquels nous avons groupé les critères ergonomiques.

N°	Questions	toujours	souvent	quelque-fois	jamais
A	**Facilité d'apprentissage**				
1	Est-ce que l'entrée des données est guidée				
2	Pour chaque champ est fourni un label				
3	les termes employés sont ils familiers aux utilisateurs et relatifs à la tâche à réaliser				
4	Est-ce que chaque libellé est facile à comprendre				
5	Est-ce que l'utilisation des images rend l'interface plus compréhensible				

6	L'organisation des items lorsque plusieurs options sont présentées est elle logique				
7	Est-il facile de trouver l'information désirée				
C	**Cohérence**				
1	Les couleurs sont elles utilisées de façon cohérente				
2	Y a-t-il localisation similaire des titres des interfaces				
3	le format des champs d'entrée de données est il toujours le même				
4	Les libellés, les noms et les abréviations sont cohérents				
E	**Efficacité d'utilisation**				
1	Existe-t-il des messages qui informent sur les tâches effectuées				
2	Est-ce que le curseur facilement repérable				
3	Y a-t-il de possibilité de modifier les commandes lors de leurs saisies				
4	Existe-t-il des accès directs vers l'information désirée				
5	Lors d'une saisie de données chaque action est elle claire				
6	Le retour d'information est il suffisant suite à une action				
7	les données qui peuvent être calculées à partir de celles saisies sont elles automatiquement faites				
8	Le temps d'exécution est il rapide				
P	**Prévention et correction d'erreurs**				
1	Il n'est pas demandé aux utilisateurs de saisir les données d'une liste				
2	Suite à une action erronée y a-t-il des messages d'erreurs				
3	Les messages d'erreurs sont ils explicites				
4	En cas d'erreur de manipulation est-il possible de revenir en arrière				

FIG. 4 – *Grille d'évaluation*

Pour évaluer le degré d'utilisabilité en se basant sur différents types d'utilisateurs nous avons construit une base de règles floues en attribuant à chaque critère d'entrée trois variables linguistiques

Facilité d'apprentissage (A) : difficile, moyenne et facile

Cohérence (C): incohérente, moyenne, et cohérente

Efficacité d'utilisation (E) : non efficace, moyenne et efficace

Correction d'erreur (P) : peu, moyen et trop

Et pour la sortie quatre variables linguistiques utilisabilité : faible, moyenne, forte et très forte.

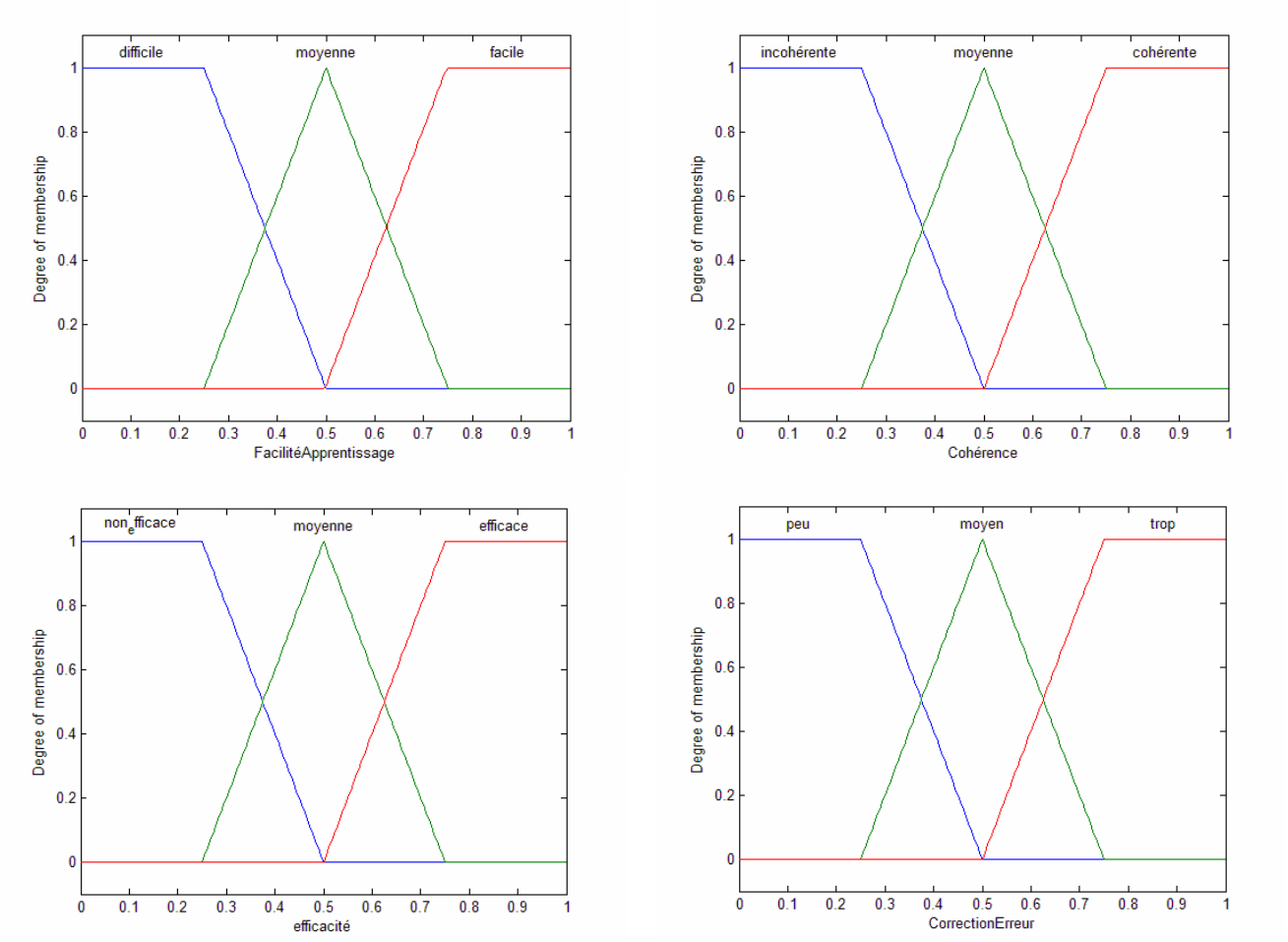


FIG. 5 – *degré d'appartenance des critères d'évaluation*

En utilisant ces variables linguistiques nous avons établie une base de règles floues formée par 81 règles.

3.2 Evaluation de l'utilité

Pour entamer cette évaluation nous avons eu recours au modèle en U [1] qui peut être adapté avec le processus d'ECD pour vérifier si les données obtenues par le système répondent aux besoins de l'utilisateur.

Ainsi l'évaluation de l'utilité consiste à confronter la tâche prescrite pour réaliser un traitement et obtenir les données intéressantes et la tâche réalisée par l'utilisateur et voir les données auquel l'utilisateur peut aboutir à partir de son interaction avec le système et la manière avec laquelle il obtient ses données. Cette évaluation est importante tout au long du

processus d'ECD lors de chaque phase pour évaluer les données obtenues après la sélection, le nettoyage et prétraitement, et la transformation.

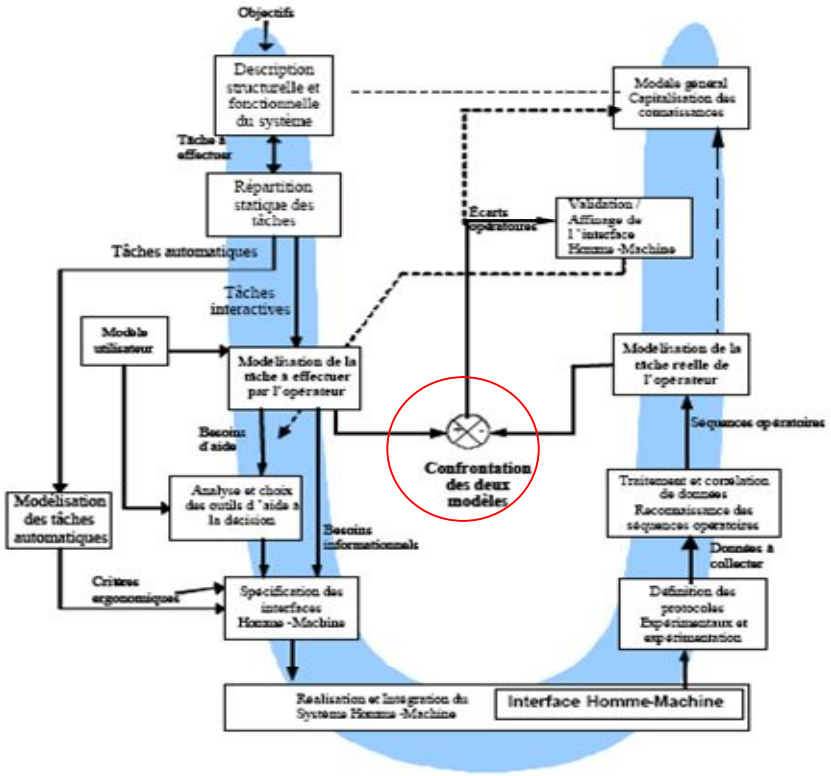


FIG. 6 – *Modèle en U [1]*

4 Test de la méthode d'évaluation

Pour l'évaluation de l'utilisabilité, comme c'est mentionné dans la précédente partie, nous avons utilisé la grille d'évaluation et la logique floue après décomposition du système en sous systèmes, comme indiqué dans la méthode « Analytic Hierararchy Process » c'est-à-dire appliquer la grille d'évaluation sur les interfaces du système. Cette méthode est appliquée sur trois différents types d'utilisateurs :

Le premier type concerne des utilisateurs possédant un peu de savoir sur le domaine sur le quel est conçu le SIAD. Alors que, le deuxième type concerne des utilisateurs n'aient pas de connaissance sur le domaine. Quand' au troisième type, il concerne les experts du domaine.

Pour le premier type d'utilisateur, nous avons sélectionné les règles actives puis calculé le poids w_i de chacune d'elles.

Les règles sélectionnées pour ce type sont :

R70: Si (A est facile) et (C est moyenne) et (E est efficace) et (P est peu) alors (U est forte)

R80: Si (A est facile) et (C est cohérente) et (E est efficace) et (P est moyenne) alors (U est tés forte)
R77: Si (A est facile) et (C est cohérente) et (E est moyenne) et (P est moyenne) alors (U est forte)
R37: Si (A est moyenne) et (C est moyenne) et (E est non efficace) et (P est peu) alors (U est faible)
R47: Si (A est moyenne) et (C est cohérente) et (E est non efficace) et (P est moyenne) alors (U est faible)
Le calcul du poids se fait par la formule

$$w_i = \mu_{Ai\,(X)} * \mu_{Ci\,(X)} * \mu_{Ei\,(X)} * \mu_{Pi\,(X)}$$

Avec $\mu_{A\,(X)}$ représente le degré d'appartenance de la facilité d'apprentissage de l'interface X et w_i est le poids de la règle active concernée.
Exemple : $w_1 = \mu_{facile\,(X)} * \mu_{moyenne\,(X)} * \mu_{efficace\,(X)} * \mu_{peu\,(X)}$

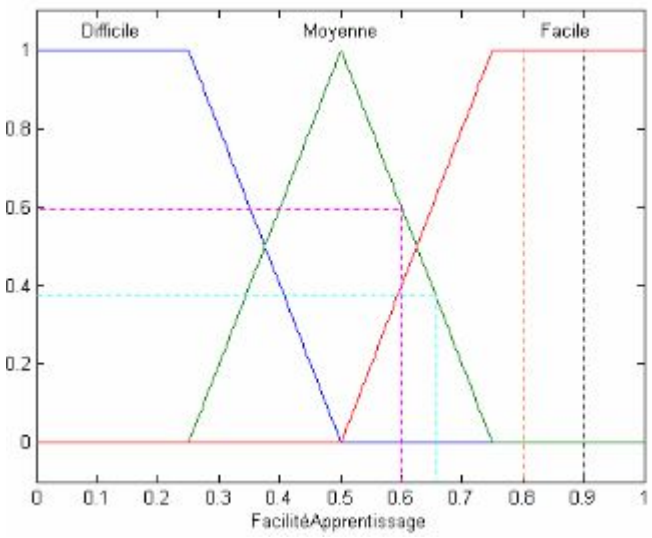
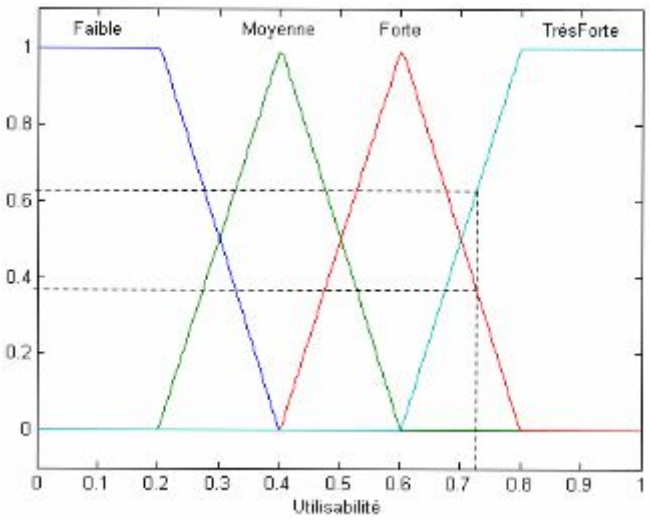


FIG. 7 – *Facilité d'apprentissage du 1er utilisateur*

Pour le 1er utilisateur nous trouvons que la valeur représentant la facilité d'apprentissage de la 1ère interface est de 0.8 et son degré d'appartenance est 1, 2 ème et la 3 ème interface est de 0.9 et son degré d'appartenance est 1, 4 ème interface est de 0.6 et son degré d'appartenance est 0.6, 5 ème interface est de 0.66 et son degré d'appartenance est 0.38 (FIG. 7)

En appliquant la méthode de la somme pondérée

$$Z = \frac{\Sigma\, w_i * x_i}{\Sigma w_i}$$



Nous avons trouvé que l'utilisabilité du système est égale à 0.73.

En projetant cette valeur sur la courbe, nous obtenons que l'utilisabilité est à 34.6% forte et à 64% très forte pour le 1er type d'utilisateur qui possède un peu de savoir sur le domaine hospitalier.

On procède de la même manière pour un deuxième type d'utilisateur, nous trouvons que le système est à 17% d'utilisabilité forte et à 88% d'utilisabilité très forte.

Finalement pour un troisième type, le système est à 14.6% d'utilisabilité forte et 86.6% d'utilisabilité très forte.

5 Conclusion

Les nouveaux systèmes informatiques utilisent des bases de données stockant de nombreuses données et ces systèmes sont conçus pour l'aide à la décision.

Pour extraire ces connaissances, l'utilisation d'un système interactif d'aide à la décision basé sur le processus d'ECD est fondamentale et pour garantir le bon fonctionnement de ce système il doit être vérifié et validé.

En effet, la validation d'un système se fait par l'évaluation des deux critères principaux qui sont l'utilité et l'utilisabilité est faite au cours de sa réalisation il s'agit de l'évaluation à priori et cela est fait dans ce mémoire par l'utilisation du Processus Unifié provenant du génie logiciel qui consiste à générer des modèles et prototypes d'interface et les améliorer par la suite tout en tenant compte de l'utilisateur car dans le cadre de travail l'évaluation est centrée utilisateur.

Références

[1]Abed M., "Méthodes et Modèles formels et semi-formels de conception et évaluation des systèmes homme-machine". Mémoire d'HDR, Université de Valenciennes et du Hainaut Cambrésis, 2001.

[2]Han J., Kamber M., "Data Mining Concepts and Techniques", Morgan Kaufmann, 2000

[3]Jacobson I., Booch G. et Rumbaugh J., "Le processus Unifié de Développement logiciel". Eds Eyrolles, Paris, 1999

[4] Kolski C**.,** "Ingénierie des Systèmes Homme Machine", 1997

[5]Larman C., "UML 2 et les Design Patterns", Eds Pearson Education, 2005

[6]Marc Nectoux, Henning Bay-Nielsen, Birth Frimodt-Moller, Robert Bauer, Jean Pierre Darlot, claude Mugnier, "Développement d'outils de data mining et d'aide à la décision", Octobre 2000

[7]René L., Gilles V. "Data mining", édition Eyrolles, 2001

[8]Scapin L., Christian J.M., "Ergonomic criteria for the evaluation of Human computer interface"

[9] Scapin D., "Guide ergonomique de conception des interfaces homme machine" : une revue de la littérature, 1986. Rapports Tachnique 77, INRIA.

[10]Senach B. "Evaluation ergonomique des interfaces homme-machine": une revue de la littérature. Rapport n°1180, INRIA, 1990.

Summary

In this article we propose a based user method to evaluate a decision support system (DSS) based on a decisional tool that is data mining process in which we consider the metric «utility» and «usability».this method aims to integrate the evaluation of interface made and data during this process to have utile data that allow knowledge's extraction.

Mesures de qualité de clustering de documents : Prise en compte de la distribution des mots clés

M. GHRIBI[1], P. CUXAC[1], J.C. LAMIREL[2], A. LELU[2] [3]

[1]INIST-CNRS, 2 allée du Parc de Brabois, 54500-Vandœuvre-lès-Nancy, France.
maha.ghribi@inist.fr ; pascal.cuxac@inist.fr
[2]LORIA Campus Scientifique BP 239 - 54506 Vandœuvre-lès-Nancy, France.
alain.lelu@loria.fr ; jean-charles.lamirel@loria.fr
[3]LASELDI / Université de Franche-Comté, 30 rue Mégevand – 25030 Besançon, France.

Résumé : Nos travaux sur une nouvelle méthode de classification non supervisée (Germen) nous ont amenés à nous interroger sur la qualité des résultats obtenus. Le problème est d'estimer si une méthode de clustering est 'meilleure' qu'une autre pour le type de données que nous traitons (données textuelles). Dans un premier temps, après avoir fait un état de l'art des méthodes existantes, nous avons appliqué quelques indices de qualité aux résultats de clustering issus de notre algorithme Germen ainsi que d'autres algorithmes communément utilisés. Ces indices de qualité ne permettant pas de sélectionner la meilleure partition, nous avons développé une nouvelle série d'indices basés sur la distribution des mots-clés. Nous présentons et discutons les résultats obtenus ainsi que les réflexions engagées pour faire évoluer l'évaluation de classifications non supervisées sur des textes.

1 Introduction

Evaluer les performances d'un algorithme de clustering (classification non supervisée) n'est pas chose aisée. Une première façon est de faire une évaluation supervisée : on compare le résultat obtenu à une référence [Yosr et Sinaoui 2009] (on peut pour cela utiliser une classification préexistante ou des corpus de référence). En ce qui concerne les corpus de tests exploitables dans le cadre de l'évaluation des méthodes de classifications, outre les corpus de test numériques classiques utilisés pour les méthodes supervisées tels qu''Iris' ou 'Mushroom', des corpus de textes indexés issus de dépêches 'Reuters' sont régulièrement utilisés dans différentes campagnes d'évaluation. Dans ce dernier cas, notre expérience montre cependant que l'indexation proposée par [Lewis et al 2004] est peu adaptée à l'évaluation de méthodes de classifications non supervisées de données textuelles [Cuxac et al. 2009].

Nous proposons de mesurer la qualité de méthodes de clustering que nous avons développées en les comparant à des méthodes choisies dans la littérature. Pour cela nous avons sélectionné quelques indicateurs de qualité communément utilisés ; leur application à nos résultats nous a amenés à une réflexion sur ces indicateurs et à faire de nouvelles propositions.

2 Les données

Nous avons utilisé un corpus bibliographique test issu de la base PASCAL, édité par L'INIST (CNRS, www.inist.fr). Ce corpus rassemble 1920 notices sur le thème de « la recherche en Lorraine », signalant des documents de type article, congrès ou thèse. Pour réaliser les classifications, nous utilisons les termes d'indexation manuelle présents dans les notices. Le corpus est indexé par 3557 descripteurs de fréquence supérieure à 1. La figure 1 illustre quelques champs d'une notice bibliographique de ce corpus.

ET : Seismic and geotechnical investigations following a rockburst in a complex French mining district
AU : DRIAD-LEBEAU (L.); LAHAIE (F.); AL HEIB (M.); JOSIEN (J. P.); BIGARRE (P.); NOIREL (J. F.); HOWER (James C.); GREB (Stephen F.)
AF : Institut National de l'Environnement Industriel et des Risques (INERIS)-Ecole des Mines, Parc Saurupt/54042 Nancy/France (1 aut., 2 aut., 3 aut., 5 aut.); GEODERIS/Metz/France (4 aut.); Charbonnages de France/Merlebach/France (6 aut.); University of Kentucky Center for Applied Energy Research/Lexington, KY/Etats-Unis (1 aut.); Kentucky Geological Survey, University of Kentucky/Lexington, KY/Etats-Unis (2 aut.)
SO : International journal of coal geology; ISSN 0166-5162; Pays-Bas; Da. 2005; Vol. 64; No. 1-2; Pp. 66-78.
EA : This paper presents the results of seismic and geotechnical studies carried out after a fatal accident that occurred during mining of the Frieda5 coal seam at Merlebach mine of HBL (Houillères du Bassin Lorrain, East France). On June 21, 2001, a violent rockburst (local magnitude of 3.6) affected the Frieda5 seam at depth of approximately 1250 m …
ED : mining; coal seams; coal mines; stress; joints; sandstone; channels; rock bursts; seismic methods; safety; risk assessment; acoustical methods; France; Lorraine; Moselle France
EG : clastic rocks; sedimentary rocks; Western Europe; Europe

FIG. 1- *Une notice bibliographique de la base PASCAL (ET = titre ; AU = auteurs ; AF = affiliations ; SO = source ; EA= résumé ; ED+EG = descripteurs)*

Un tel corpus est ensuite transformé en une matrice booléenne documents×mots, très creuse par nature. Bien que ces données ne proviennent pas directement de textes, nous avons constaté que la répartition des termes avait l'allure zipfienne (relation linéaire entre le log des fréquences de termes et le log de leur rang) caractéristique des données textuelles. Ce corpus peut-être obtenu en contactant les auteurs.

3 Les méthodes de clustering et leurs résultats

3.1 Méthodes de clustering utilisées

Dans cette partie, nous allons décrire les méthodes de clustering utilisées : des méthodes neuronales (SOM, IGNG, K-means Axiales, Analyse en Composantes locales++), et des méthodes opérant sur des graphes (Walktrap, Germen).

3.1.1 Méthodes neuronales

Nous nous sommes intéressés à quatre méthodes qui sont les K-means Axiales (KMA), l'Analyse en Composantes Locales++ (ACL++), SOM (carte auto-organisatrice de Kohonen) et IGNG (Incremental Growing Neural Gas)

Dans **la méthode KMA** (K-Means Axiales) [Lelu 1994], les vecteurs-documents sont normalisés selon la métrique de Hellinger, particulièrement adaptée aux données textuelles, et sont affectés à des axes de classe (« vecteurs-neurones ») pointant vers les zones de forte

densité des données, avec des degrés de centralité dans leur classe plus ou moins prononcés selon le principe des centres mobiles des K-means.

La méthode ACL++ (Analyse en composantes locales dans l'espace sphéré des données) : L'Analyse en composantes locales (Lelu 1994) est comme Germen, décrite plus loin, une méthode de détermination d'axes de clusters par montée en gradient de vecteurs-neurones, sur le paysage de densité des données. La densité est déterminée en tout point de cet espace à partir d'un voisinage de rayon fixe, à la différence de Germen où la densité est adaptative.

La variante ACL++, à la différence de l'ACL, n'est pas réalisée dans l'espace des descripteurs, mais dans l'espace des K premiers vecteurs singuliers de la matrice des données, où K est déterminé par un test statistique [Lelu et Cadot,2010]. Elle aboutit à une "normalisation" des densités dans cet espace et à une adaptativité aux différences de densité, de façon moins locale que dans Germen.

La méthode SOM [Kohonen,1982] est une méthode où les vecteurs-neurones s'auto-organisent au fil des données sous forme d'une structure de voisinage bien définie (« ficelle » unidimensionnelle, grille bidimensionnelle ou multidimensionnelle). Considérée comme une méthode statique, c'est-à-dire à nombre de neurones pré-fixé, son algorithme débute par l'initialisation de la carte de voisinage avec une sélection aléatoire des neurones. A chaque itération, l'insertion d'une nouvelle donnée induit un auto-ajustement de la carte par la modification du vecteur de référence du neurone le plus proche de la donnée d'entrée ainsi que ses voisins directs.

La méthode IGNG [Prudent, Ennaji,2004], [Prudent, Ennaji,2005a], [Prudent, Ennaji,2005b] est une méthode neuronale très différente de SOM puisque d'une part, elle n'impose aucune structure de voisinage et d'autre part, le nombre de neurones varie au cours de l'apprentissage. Il y a possibilité de suppression et de création de neurones. La suppression est liée à l'introduction de la notion d'âge aux liens entre les neurones. Si à une itération l'âge d'un lien atteint la maturité, celui-ci est supprimé. Les neurones se trouvant isolés sont automatiquement supprimés. La création se fait à chaque itération si certaines conditions sont vérifiées. En effet, avec l'introduction d'une nouvelle donnée, et avant d'effectuer l'apprentissage, on vérifie si la création d'une nouvelle classe est nécessaire. La règle est la suivante : s'il existe au moins deux neurones dont la distance par rapport à la nouvelle donnée est inférieure à une certaine valeur σ préfixée, la création du neurone est inutile. Sinon, la nouvelle donnée représente le nouveau neurone. σ est calculé à partir de toutes les données présentes dans l'échantillon. C'est la distance moyenne qui sépare toutes les données de la donnée centrale.

Une caractéristique très intéressante de IGNG, en dehors de son aspect dynamique, est sa tolérance aux données bruitées. En effet, l'algorithme affecte à chaque neurone un âge depuis sa création. Si à une certaine itération, un neurone atteint l'âge de maturité, il passe de l'état d'un neurone embryon à un neurone mature. L'âge de maturité correspond donc au nombre minimal d'activation pour qu'un neurone ne soit plus considéré comme résultant de données bruitées.

3.1.2 Méthodes opérant sur des graphes

Parmi les méthodes qui opèrent sur des graphes, nous nous somme intéressés à deux méthodes, Walktrap et Germen.

Walktrap [Pons, Latapy : 2006] ou méthode des marches aléatoires (Random Walks) a pour but de décomposer le graphe en un certain nombre de « communautés » ou classes. Son principe est similaire à celui de la classification ascendante hiérarchique. En effet, commençant par une partition où chaque donnée forme une classe, à chaque paquet de t itérations de marches aléatoires, elle fait fusionner les deux classes qui d'une part, présentent au moins un lien entre leurs données et d'autre part, en se basant sur la méthode de Ward, minimisent la moyenne de la distance au carré de chaque sommet à sa communauté. La différence de Walktrap par rapport à la méthode ascendante hiérarchique est qu'elle calcule la distance entre les nœuds du graphe à partir de la matrice d'adjacence. Cette dernière permet de déterminer la matrice de transition entre les éléments du graphe (nœuds et communautés). L'algorithme se termine par une partition qui contient une seule classe regroupant tous les nœuds. Le choix de la meilleure partition est fait de façon à ce quelle maximise le critère de « Modularité » [Newman, Girman, 2004] décrit plus bas.

Germen [Cuxac, et al 2009], [Lelu et al. 2006] est une méthode de clustering de graphes. Elle se base sur la notion de densité du nuage des vecteurs-données. Son principe se résume en la détection des maxima de densité à chaque itération et en la prise en compte des perturbations locales dues à l'introduction d'un nouvel élément. A chaque élément du graphe (document) on attribue une densité. Le partitionnement du graphe se fait par le repérage des nœuds les plus denses. Ces derniers représentent les « chefs de classes ». Leurs zones d'influence s'élargissent au fur et à mesure par rattachement unique ou partagé de leurs voisins de plus en plus éloignés. Ce rattachement se base sur la notion d'héritage des étiquettes des chefs de classes. Plusieurs règles peuvent exister, par exemple :

- Un nœud hérite du numéro du chef de classe de son voisin le plus surplombant (ayant une densité la plus importante que la sienne et son 1-voisinage). Si celui-ci n'existe pas, on crée une nouvelle classe. Dans ce cas, un document ne peut appartenir qu'à une seule classe (classification non recouvrante).
- Un nœud hérite des numéros de chefs de classes de tous ses voisins surplombants. Dans ce cas, un document peut appartenir à plusieurs classes (classification recouvrante).

Pour la construction du graphe, Germen peut utiliser différentes méthodes (K plus proches voisins par exemple) mais, suite à nos expériences précédentes [Cuxac et al. 2006], nous leur avons préféré ici la méthode « Tournebool » [Cadot, 2006]. C'est un algorithme de validation statistique des liens entre vecteurs-données qui consiste en la génération d'un grand nombre de matrices booléennes de sommes marginales équivalentes à celles de la matrice des données. Pour chaque couple de documents on calcule son support (nombre de mots clés commun) dans chaque matrice générée. On obtient ainsi une distribution des supports des deux documents. Si le support dans la matrice de données initiale est supérieur à un certain seuil prédéfini dans la distribution, le lien est considéré non dû au hasard, donc valide.

3.2 Résultats de clustering

Pour toutes les méthodes on a fait varier les paramètres pour avoir plusieurs partitions avec des nombres de classes différents, par exemple le nombre de classes pour les méthodes qui le considèrent comme paramètre, comme SOM et KMA. Alors qu'ACL++ fixe le nombre de classes à la seule « dimension structurelle » déterminée par son test statistique, ici autour de 155, on a fait varier le sigma d'IGNG. Avec Germen, on a seuillé les liens, ce qui a permis d'obtenir des partitions différentes. Avec Walktrap, on a varié le nombre d'itérations.

Méthode	Nombre de Classe	Nbr de docs classés	Taille de la plus grosse classe	Nbr de classes de taille < 10	Nbr de classes de taille > 100
SOM	49	1920	138	21	3
Germen	61	1707	450	46	4
Walktrap	67	1624	309	46	5
KMA	155	1920	108	77	1
ACL++	151	1920	52	78	0
IGNG	169	1920	338	110	1

TAB. 1 – *Résultats de classification des différentes méthodes de clustering*

Le tableau 1 montre pour chaque méthode la partition qui peut être considérée comme la meilleure en se basant sur les indices de qualité présentés par la suite. Pour les méthodes Germen et Walktrap, le nombre de documents classés est inférieur à la taille de l'échantillon utilisé : la construction du graphe a isolé certains documents qui ne sont pas considérés dans la classification. On remarque un comportement équivalent entre ces deux méthodes en termes de nombre de classes et taille des classes. Avec la présence de très grosses classes, on attend une hétérogénéité à l'intérieur de ces classes. Les méthodes KMA, ACL++ et IGNG présentent un nombre très important de classes de faible taille. Ceci risque d'influencer la répartition des similarités entre les classes. SOM avec une partition de 49 classes présente un équilibre de point de vue structure des classes (absence de très grosses classes et faible proportion de classes de petite taille). ACL++ également.

4 Les indices de qualité : un état de l'art

On distinguera dans cette brève revue deux familles d'indices de qualité de clustering : les indices opérant sur des graphes et ceux basés sur les distances.

4.1 Indices de qualité opérant sur les graphes

Les indices de qualité opérant sur des graphes s'intéressent aux liens entre les nœuds à l'intérieur des graphes. Ils se basent sur le principe que les nœuds appartenant à une même classe sont plus liés entre eux qu'avec les points appartenant à des classes différentes. Plusieurs formalisations ont été développées : la « Performance » et la « Modularité » utilisent des critères inter et intra classes afin de mesurer à quel point les classes sont formées par des éléments homogènes, et les classes bien séparées.

Dans cette partie, on note G(E,V) un graphe où E représente la liste des *n* nœuds et V représente la liste des arêtes dans le graphe. On note aussi $P=(C_1,\ldots,C_k)$ une partition du graphe.

La « Performance » [Van Dongen, 2000] introduit la notion de couple de nœuds correctement interprétés qui désigne à la fois les couples de nœuds liés appartenant à une même classe et les couples de nœuds non liés appartenant à deux classes différentes. La Performance calcule la fraction des couples de nœuds correctement interprétés par rapport au nombre total de couples de nœuds : $Performance(P) = \frac{m(P) + \sum_{\{u,v\} \notin E, u \in C_i, v \in C_j, i \neq j} 1}{\frac{1}{2} n(n-1)}$

$m(P)$ représente le nombre de liens à l'intérieur des classes.

Une augmentation de la Performance signifie que d'une part, les classes sont bien séparées et d'autre part, que les nœuds à l'intérieur des classes sont bien liés entre eux. Par conséquent, plus la performance est proche de 1 meilleure est la classification.

La « Modularité » [Newman, Girman ,2004] calcule, à coté de la proportion des liens intra classes, la proportion des liens inter classes.

$$Q(P) = \sum_{C \in P} (e_C - a_C^2)$$

e_C représente la fraction des liens Intra classe : $e_C = \frac{\sum_{u,v \in C, \{u,v\} \in E} 1}{|E|}$

a_C représente la fraction des liens qui ont au moins une extrémité dans C

$$a_C = \frac{\sum_{u \in C, v \in E, \{u,v\} \in V} 1}{|E|}$$

a_C fait intervenir les relations que possède la classe C avec les autres classes (les liens inter-classes).

Plus la proportion des liens intra-classe augmentent et les liens inter-classes diminuent, plus les documents à l'intérieur des classes sont liés et les classes sont séparées entre elles. Donc la meilleure partition c'est celle qui maximise la modularité.

4.2 Indices de qualité basés sur la distance

Les indices inertiels [Lebart et al, 1982] sont les plus connus et les plus utilisés pour évaluer la qualité d'une classification.

- L'inertie intra-classes permet de mesurer le degré d'homogénéité entre les objets appartenant à la même classe. Elle calcule leurs distances par rapport au point représentant le profil de la classe.

$$Intra = \frac{1}{n} \sum_{C \in P} \frac{1}{2n_c} \sum_{i \in C} \sum_{j \in C} d(i, j)^2$$

- L'inertie inter-classes mesure le degré d'hétérogénéité entre les classes. Elle calcule les distances entre les points représentant les profils des différentes classes de la partition.

$$Inter = \frac{1}{n} \sum_{C \in P} n_C d^2(c, c_G)$$

Avec c le centre de la classe C et c_G est le centre du nuage de points.

Plus les données à l'intérieur des classes sont homogènes, plus leurs distances par rapport au point représentant la classe sont faibles. Par conséquent, une valeur faible de l'inertie intra-classes décrit une homogénéité des données à l'intérieur des classes.

Plus les classes sont hétérogènes entre elles, plus les distances entres les points représentants les profils des classes sont élevées. Donc, une valeur élevée de l'inertie inter-classes traduit une hétérogénéité entre les classes. Cet indice a le défaut d'augmenter quand on augmente le nombre de classes.

Plusieurs autres indices de qualité qui utilisent la distance entre les individus ont été développés dont l'indice de Dunn, l'indice de validation de Davies–Bouldin et la Silhouette.

Les indices de Dunn et de Davies-Bouldin mélangent à la fois les inerties Intra classes et les inerties Inter classes.

L'indice de Dunn [Dunn,1974] est décrit par la formule suivante :

$$Dunn = \min_{1\le i\le n}\left\{\min_{\substack{1\le j\le n\\ j\neq i}}\left\{\frac{d(c_i,c_j)}{\max_{1\le k\le n}(d'(c_k))}\right\}\right\}$$

Il cherche la distance minimale qui sépare deux classes dans la partition tout en tenant compte de la distribution des éléments à l'intérieur des classes. Plus cette distance est grande meilleure est la partition.

L'indice de Davies-Bouldin (DB) [Davies et Bouldin , 2000] traite chaque classe individuellement et cherche à mesurer à quel point elle est similaire à la classe qui lui est la plus proche. L'indice DB est formulé de la façon suivante :

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{j\neq i}\left\{\frac{I(c_i)+I(c_j)}{I(c_i,c_j)}\right\}$$

Pour chaque classe i de la partition, on cherche la classe j qui maximise l' « indice de similarité » décrit comme suit $R_{ij} = \frac{I(c_i)+I(c_j)}{I(c_i,c_j)}$

$I(c_i)$ représente la moyenne des distances entre les documents appartenant à la classe C_i et son centre. Et $I(c_i,c_j)$ représente la distance entre les centres des deux classes C_i et C_j.

La meilleure partition est donc celle qui minimise la moyenne de la valeur calculée pour chaque classe. En d'autres termes, la meilleure partition est celle qui minimise la similarité entre les classes.

L'indice Silhouette [Rousseeuw, 1987] est différent, des indices de qualité traités ci-dessus ; il travaille à l'échelle microscopique, c'est à dire qu'il s'intéresse aux documents en particulier et non pas aux classes. Le but de Silhouette est de vérifier si chaque document a été bien classé. Pour cela, et pour chaque document i de la partition, on calcule la valeur suivante :

$$-1 \le S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \le 1$$

a(i) représente la distance moyenne qui le sépare des autres documents de la classe à laquelle il appartient et *b(i)* représente la distance moyenne qui le sépare des documents appartenant à la classe la plus proche.

Quand *S(i)* est proche de 1, le document est bien classé : la distance qui le sépare de la classe la plus proche est très supérieure à celle qui le sépare de sa classe. Par contre, si S(i) est proche de -1, cela veut dire que le document est mal classé. Mais si S(i) est proche de 0 alors il pourrait également être classé dans la classe la plus proche.

L'indice Silhouette de la partition est calculé à partir de la moyenne entre les indices de ses éléments.

Les mesures de qualité exposées ci-dessus sont intéressantes en soi, mais le domaine du traitement des données textuelles est un peu spécifique puisque les données ne sont pas décrites par des variables quantitatives mais plutôt par les descripteurs qualitatifs que sont les mots-clés caractérisant chaque document. Il n'est donc pas immédiat de définir les liens entre les documents afin d'utiliser les mesures de qualité qui se basent sur des graphes. De plus, le calcul des distances utilisables par les indices inertiels, les indices de Dunn, de Davies–

Bouldin et Silhouette sont délicats. En effet, les profils des documents sont des vecteurs binaires très creux dans un espace des mots-clés généralement de très grande dimension.

4.3 Les résultats

Nous allons maintenant appliquer les indices de qualité décrits précédemment sur les nos résultats de clustering. Pour cela, nous avons utilisé la distance de Jaccard pour calculer la distance entre les documents [Jaccard 1901]. Cette distance est adaptée aux données de profils binaires. Son principe est le suivant : pour chaque couple de documents D_i et D_j on a le tableau suivant :

$D_i \setminus D_j$	1	0
1	a	b
0	c	d

TAB. 2 – *Tableau croisé entre deux documents permettant de calculer une vaste famille d'indices de (dis-)similarité, en particulier leur distance de Jaccard*

a représente le nombre de descripteurs commun entre les deux documents, b représente le nombre de descripteurs présents dans *Di* et non pas dans D_i, c représente le nombre de descripteurs présents dans D_i et non pas dans D_i et d représente le nombre de descripteurs qui sont absents dans les deux documents.

La distance de Jaccard prend la forme suivante : $d(Di, Dj) = \frac{b+c}{a+b+c}$

Plus le nombre de descripteurs communs entre les documents est élevé plus la distance est faible.

Pour les méthodes de clustering opérant sur des graphes (Walktrap et Germen), on a construit le graphe avec la méthode de Tournebool décrite en 3.1.2 [Cadot, 2006].

Examinons maintenant le comportement des indices de qualité sur nos données documentaires. La Performance et l'inertie intra-classes illustrées dans la figure 2 montrent une forte dépendance avec le nombre de classes dans la partition. Selon ces deux indices, toutes les méthodes sont équivalentes et le choix de la meilleure partition devient difficile.

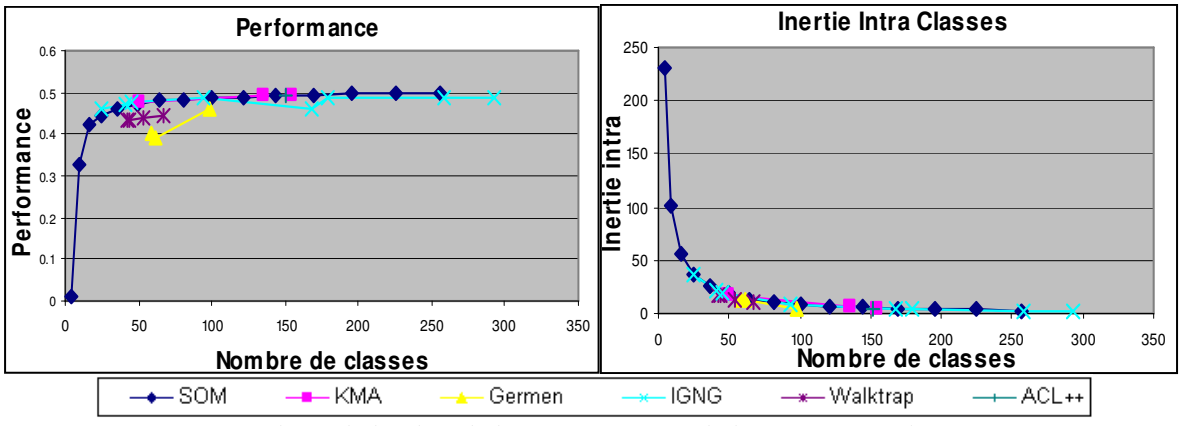


FIG. 2 – *Les valeurs de l'indice de la Performance et de l'Inertie Intra-Classes en fonction du nombre de classes produites par les différentes méthodes de clustering*

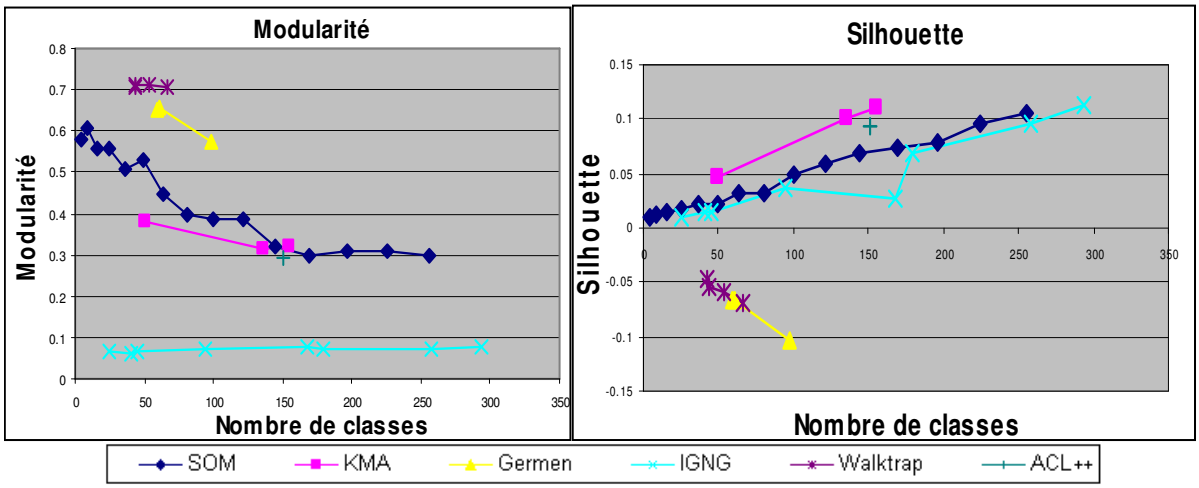


FIG. 3 – *Les valeurs de l'indice de Modularité et de Silhouette en fonction du nombre de classes produites par les différentes méthodes de clustering*

La modularité présente des valeurs très différentes selon les méthodes utilisées (figure 3). Cependant, on remarque que les valeurs associées à Germen et Walktrap sont les plus élevées. Ceci est attendu car elles présentent de très grosses classes et donc le nombre de liens intra classes est très important. Mais les résultats de la Modularité ne sont pas satisfaisants puisqu'ils n'ont pas permis de détecter l'hétérogénéité à l'intérieur des classes produites par ces deux méthodes.

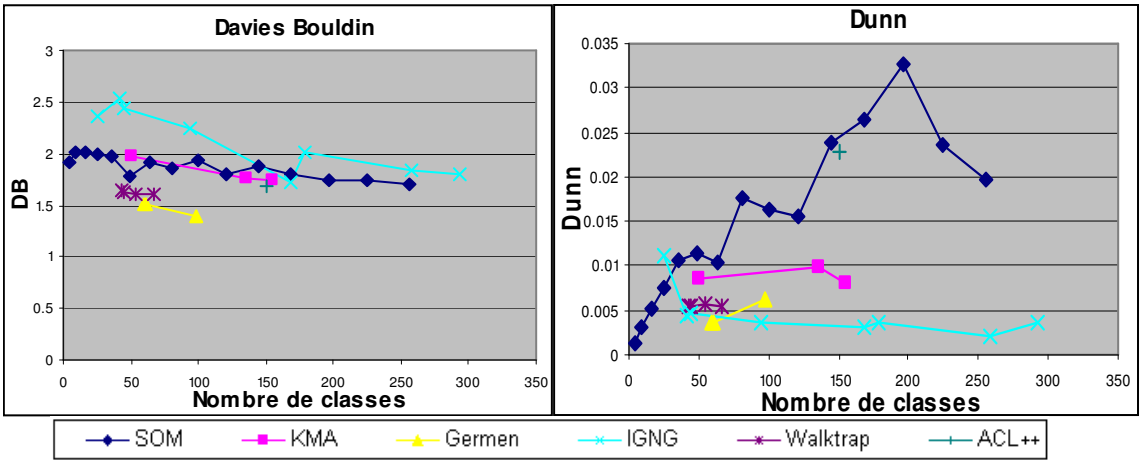


FIG. 4– *Les valeurs de l'indice de Davies Bouldin et de Dunn en fonction du nombre de classes produites par les différentes méthodes de clustering*

L'indice Silhouette illustré dans la figure 3 montre des valeurs absolues très proches de zéro. On peut conclure que les classes sont très proches les unes des autres et que les documents présentent une confusion au niveau de leur classification. Ce que l'on peut reprocher à Silhouette est que les documents appartenant à des classes caractérisées par une

grande dispersion de leurs éléments vont avoir des valeurs négatives s'ils sont proches de classes de faibles tailles. Ceci est clairement visible avec les partitions de Germen et Walktrap qui présentent à la fois des grosses classes et des classes de faible taille.

L'indice de Davies-Bouldin illustré dans la figure 4 montre une équivalence entre les résultats des deux familles de méthodes de clustering. Les valeurs sont très proches. Les méthodes neuronales présentent un comportement similaire de même que les méthodes Walktrap et Germen. L'indice de Dunn présente de très faibles variations avec des valeurs proches de zéro (figure 4).

5 De nouveaux indices de qualité : mesures basées sur les distributions des descripteurs

5.1 Rappel, précision, F-mesure

Les mesures de qualité habituellement utilisées ne sont donc pas optimales pour évaluer les résultats de classification non supervisée sur des corpus de textes. Notre démarche a alors consisté à développer des indices qui tiendraient mieux compte du type de nos données et qui seraient indépendants de la méthode de clustering utilisée, tout cela sans classification de référence.

La notion de Rappel, Précision et F-mesure a été introduite par Van Rijsbergen [Van Rijsbergen, 1979]. Elle se base sur le fait qu'un système de recherche documentaire est efficace s'il permet de restituer un maximum d'informations pertinentes. La Précision (P) détermine le pourcentage de documents pertinents restitués pour une requête donnée et le Rappel (R) calcule le pourcentage de documents pertinents restitués par rapport au nombre total des documents pertinents pour cette même requête. Par conséquent, le système de recherche d'information est efficace quand le Rappel et la Précision sont proches de 1. La F-mesure est la moyenne harmonique du Rappel et de la Précision : $F_{mesure} = 2\left[\frac{1}{R} + \frac{1}{P}\right]$

Il est clair que ces définitions de Rappel Précision F-mesure reposent sur des connaissances préalables sur la nature des documents (pertinente ou pas). Ce qui rend ces indices inapplicables dans le cas d'une classification non supervisée à cause de l'absence de classification de référence.

Ces indices ont cependant été adaptés au cas du clustering non supervisé [Lamirel et al, 2003]. Les mesures ne se font plus sur les documents mais sur les mots clés. L'idée est de mesurer l'homogénéité des classes en étudiant la distribution des mots clés à l'intérieur des classes. On introduit la notion de « mots propres » aux classes : chaque classe est caractérisée par un ensemble de mots clés dont les poids à l'intérieur de la classe par rapport à leurs poids dans la partition sont maximaux.

Plus explicitement, pour une partition $P = (C_1,...,C_k)$, on définit pour chaque classe C_i l'ensemble des mots propres suivant :

$$S_C = \left\{ p \in d, d \in C_i C \;\middle|\; \overline{W}_C^p = \max_{C' \in P}(\overline{W}_{C'}^p) \right\} \text{ avec } \overline{W}_C^p = \frac{\sum_{d \in C} W_d^p}{\sum_{C' \in P}\sum_{d \in C'} W_d^p}$$

où W_p^d représente le poids de la propriété *p* pour un document *d* et $\overline{W}_C^p$ représente le rapport du poids cumulé de la propriété *p* dans la classe C à son poids total dans la partition. On définit aussi l'ensemble des classes propres comme suit : $\overline{P} = \{C \in P | S_C \neq \varnothing\}$

A partir de ces propriétés propres, les valeurs globales de Rappel et de Précision sont calculées de la manière suivante :

$$R = \frac{1}{|\overline{P}|} \sum_{C \in \overline{P}} \frac{1}{S_C} \sum_{p \in S_C} \frac{|c_p|}{|P_p|} \; ; \; P = \frac{1}{|\overline{P}|} \sum_{C \in \overline{P}} \frac{1}{S_C} \sum_{p \in S_C} \frac{|c_p|}{|c|}$$

c_p représente l'ensemble des documents de la classe C possédant la propriété *p* et P_p représente l'ensemble des documents de la partition *P* possédant la propriété *p*

Le Rappel mesure l'exhaustivité du contenu des classes. Il calcule, pour chaque mot propre associé à une classe donnée, la proportion des documents appartenant à la même classe et dans lesquels il apparaît par rapport au nombre total de documents contenant ce mot. Plus les classes présentent des mots propres exclusifs, plus la valeur globale du Rappel augmente corrélativement à la qualité du clustering.

La Précision permet de mesurer l'homogénéité du contenu des classes générées. Elle calcule pour un mot propre associé à une classe donnée, la proportion dans la classe des documents qui contiennent ce mot. Si les documents appartenant à une même classe ont tendance à avoir des descripteurs similaires, la précision globale augmente corrélativement à la qualité du clustering.

La F-mesure est définie de manière équivalente à celle de C.J.Van Rijsbergen.

Les valeurs globales de Rappel Précision et F-mesure varient entre 0 et 1. La qualité du clustering est meilleure quand R, P et F-mesure sont proches de 1. Cependant, le comportement de R et de P sont différents vis à vis du nombre de classes. En effet, plus le nombre de classes augmente, plus les effectifs à l'intérieur des classes diminuent, plus les valeurs $(|c_p| / |P_p|)$ diminuent et les valeurs $(|c_p| / |c|)$ augmentent. Par conséquent, R diminue avec le nombre de classe et P augmente avec le nombre de classes. Dans ce cas, un compromis possible entre R et P est de choisir la partition qui correspond à l'écart entre le Rappel et la Précision le plus faible. Le cas idéal est celui qui correspond à la coïncidence entre les deux valeurs.

5.2 Les résultats

Visualisons maintenant le comportement des indices Rappel Précision et F-mesure sur les résultats des méthodes de clustering (figure 5). Nous remarquons que les valeurs calculées ne présentent plus une similarité comme les autres indices de qualité traités ci-dessus. En effet, l'écart entre les méthodes se creuse aussi bien avec le Rappel qu'avec la Précision. Bien que ces indices aient des comportements monotones avec le nombre de classes (le Rappel diminue et la Précision augmente avec le nombre de classes), leur dépendance n'est pas aussi forte que celle de l'Inertie Intra classe et la Performance.

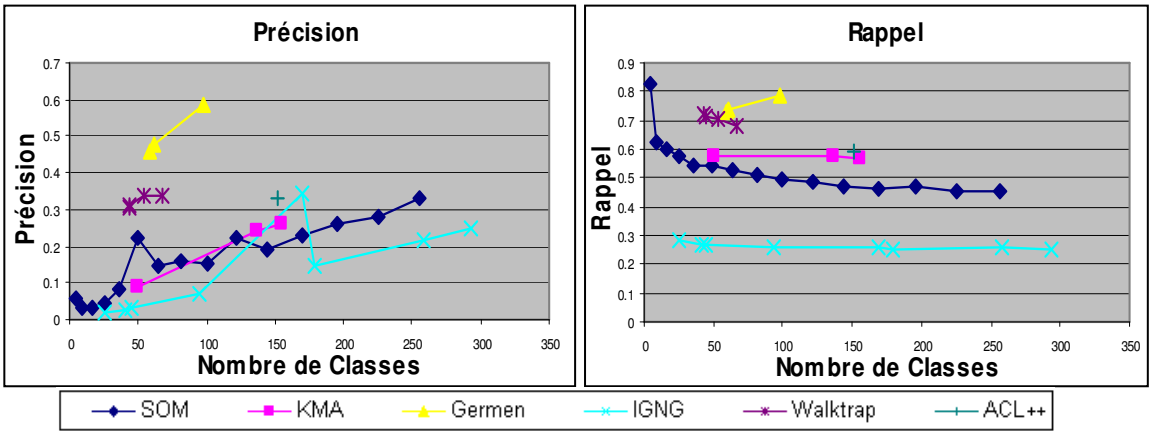


FIG. 5 – *Les valeurs des indices de Précision et Rappel en fonction du nombre de classes calculées pour les différentes méthodes de clustering*

On remarque aussi que Germen et Walktrap présentent des valeurs d'indices plus élevées. Les indices de Rappel Précision sont basés sur des moyennes calculées par classes. Ils se comportent comme des Macro-mesures et ne permettent donc pas d'identifier des classes hétérogènes comme celles produites par les méthodes Germen et Walktrap. Une alternative basée sur ces mesures est cependant possible. Elle consiste à calculer les valeurs de Rappel et de Précision globales en moyennant directement les valeurs de Rappel et de Précision des propriétés propres. Ces nouvelles mesures pourraient être considérées comme des Micro--mesures, indépendantes de la taille des classes.

6 Conclusions et Perspectives

L'évaluation d'un résultat de clustering reste une tâche très délicate. Des méthodes existent mais nos travaux montrent qu'elles ne sont pas satisfaisantes dans le cadre de l'analyse de données textuelles. Notre tentative d'élaborer de nouveaux indices basés sur la distribution des mots clés dans les classes n'a actuellement pas donné de bons résultats. Nous proposons d'axer nos travaux sur les trois points suivants :

- **Construction d'une classification de référence :** à partir des résultats de clustering obtenus sur notre corpus « La recherche en Lorraine », nous avons commencé la construction d'une classification « idéale » en corrigeant manuellement les erreurs constatées par des experts. Ce travail est long et fastidieux et il est vraisemblable qu'il n'y aura pas un résultat qui sera unique et indiscutable mais que l'on aura plutôt une certaine vision qui pourrait être différente en fonction des experts consultés. L'intérêt d'une telle approche est de garder un jeu de données réel, car des méthodes appliquées sur des jeux de données tests, plus ou moins simplistes, peuvent donner de bons résultats mais ne pas résister à l'épreuve des données réelles (données bruitées, incomplètes…).
- **Jeux de données artificielles :** une autre solution pourrait être de simuler artificiellement des données susceptibles d'être regroupées en un nombre déterminé

de classes. Le problème est alors de coller à la réalité de nos données particulières, à distributions des termes typiquement zipfiennes.

- **Construction d'indicateurs de qualité spécifiques :** cette alternative consiste à approfondir les approches basées sur les Micro-mesures telles que celles mentionnées à la fin de cet article. Selon nos expériences en cours, ces mesures semblent en effet présenter une alternative intéressante pour différencier les résultats de classification homogènes des résultats hétérogènes.

7 Références

Cadot M. (2006): Extraire et valider les relations complexes en sciences humaines : Statistique, motifs et règles d'association. Thèse Université de Franche-Comté, Besançon. 2006.

Cuxac P., Lelu A., Cadot M. (2009) : Suivi incrémental des évolutions dans une base d'information indexée : une boucle évaluation /correction pour le choix des algorithmes et des paramètres. 2ème conférence Internationale sur les systèmes d'informations et Intelligence Economique SIIE 2009, Hammamet Tunisie.

Davies D.L., Bouldin D.W.(2000): A cluster separation measure. IEEE Trans. Pattern Anal. Machine Intell, 1(4), 224-22.

Dunn J. (1974): Well Separated clusters and optimal fuzzy partitions. Journal of Cybernetics,4, 95-104.

Jaccard P. (1901) Bulletin de la Société Vaudoise des Sciences Naturelles 37, 241-272.

Kohonen T. (1982) : Self-Organized Formation of Topologically Feature Maps. Department of technical physics, Helsinki University of technology, Espo, Finland.

Lamirel J.C., François C., Al Shehabi S., Hoffmann M. (2003): New classification quality estimators for analysis of documentary information: Application to patent analysis and web mapping. In 9th International Conférence on Scientometrics an Informetrics - ISSI 2003, beiging, Chine.

Lebart L.,Maurineau A., Piron M. (1982): Traitement des données statistiques. Dunod, Paris.

Lelu A. (1994) : Clusters and Factors : neural algorithms for a noval representation of huge and highly multidimentionnal datasets. In New Approaches in Classification and Data Analysis, E.Diday, Y.Lechevallier al. editors, pp 241-248, Springer- Verlag, Berlin.

Lelu A., Cadot M. (2010) Slimming down a binary datatable: structural dimension and essential content. Soumis à COMPSTAT'2010, Paris, 22-27 Août 2010.

Lelu A.,Cuxac P.(2006): Classification dynamique d'un flux documentaire : une évaluation statique préalable de l'algorithme GERMEN. 8ème journée internationales d'Analyse statistique des Données textuelles, France.

Lewis D.D., Yang Y., Rose T., Li F., (2004) : RCV1: A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research, 5:361-397, 2004.

Naïja Y., Sinaoui K.B. (2009) : A novel measure for validating clustering results applied to road traffic. In Proceedings of the Third international Workshop on Knowledge Discovery From Sensor Data (Paris, France, June 28 - 28, 2009). SensorKDD '09

Newman M.E.J., Girman M. (2004) : Finding an evaluating community structure in networks.Physical Review E, 69(6).

Pons P., Latapy M. (2006) : Computing communities in large networks using random walks. Journal of Graph Algorithms and Application.

Prudent Y., Ennaji A. (2004): Clustering incrémental pour un apprentissage distribué : vers un système évolutif et robuste. In Conférence CAp 2004,

Prudent Y., Ennaji A. (2005 a): A new learning algorithm for incremental self-organizing maps. TESANN'05 proceeding, European Symposium on Artificial Neural Networks, Brugs, Belgium,

Prudent Y., Ennaji A. (2005 b): An Incremental Growing Neural Gas learns Topologies. Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International

Rousseeuw P.J. (1987): Silhouettes:a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

Van Dongen S.M. (2000) : Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht

Van Rijsbergen C.J. (1979) : Information Retrieval. Butterworths, London.

Abstract: Our work on a new method for unsupervised classification (Germen) led us to question ourselves on the quality of results. The problem is to estimate whether a clustering method is 'better' than another for text data. Initially, after a state of the art of existing methods, we applied some quality indices to clustering results from our Germen algorithm and other algorithms commonly used. These quality indices did not allow us to select the best partition, so we have developed a new series of indices based on the distribution of keywords. We present and discuss the results obtained and reflections initiated to evolve the evaluation of unsupervised text classification.

Un corpus de référence pour l'évaluation de la fouille d'opinion dans le contexte industriel du projet DOXA

Patrick Paroubek*, Martine Huraul-Plantet*, Catherine Gouttas**, Alexander Pak*

*LIMSI-CNRS
{pap,mhp,alexpak}@limsi.fr

**THALES Communications (TCF)
catherine.gouttas@fr.thalesgroup.com

Résumé. Nous présentons la construction d'un corpus de référence pour l'évaluation de systèmes de fouille d'opinion dans le cadre du projet industriel DOXA. Après avoir présenté le projet DOXA, nous faisons une présentation des différentes approches de l'analyse de sentiment et d'opinion avant de faire un retour sur les principales campagnes d'évaluation qui ont eu lieu. Nous présentons ensuite le corpus issu du WEB et concernant les jeux vidéo, qui sera utilisé pour constituer la référence ainsi que ses annotations. Avant de conclure, nous présentons l'interface d'annotation basée sur Knowtator.

1 Le projet DOXA

Les traditionnels outils de mesure de l'opinion ne sont plus adaptés aux nouveaux usages du Web 2.0, lesquels imposent d'apporter des réponses industrielles aux besoins d'analyse et de suivi de l'opinion « en temps réel ». En adressant un ensemble de problématiques liées au traitement automatique des opinions et sentiments dans des corpus de données multilingues, intégrant à la fois des données non-structurées et structurées, le projet DoXa[1] vise à apporter des éléments de réponse à ces nouveaux besoins technologiques.

Le projet vise en effet à spécifier, développer et/ou paramétrer des composants, des ressources et des chaînes de traitement qui permettront de :

- détecter automatiquement les thèmes abordées dans de grands volumes de textes rédigés en français et en anglais,
- détecter automatiquement les sentiments et opinions exprimés dans de grands volumes de textes rédigés en français et en anglais,
- mettre en relation les sentiments et opinions exprimés avec les thèmes sur lesquels ces sentiments et opinions portent,

[1] Le Consortium DoXa est composé de 12 partenaires - Pme, grands groupes et laboratoires académiques - : Arisem, IlObjects, OpinionWay, Pertimm, EDF, Meetic, Thales, ChART, IGM, LIMSI, LIP6 et LUTIN. Elda est partenaire associé du projet.

– transformer les informations extraites des textes en informations structurées, coupler ces nouvelles informations sur l'opinion aux informations caractérisants ceux qui les émettent et les media qui les contiennent, en vue d'en déduire, en s'appuyant sur des techniques d'analyse des données, des connaissances synthétisées et exploitables,
– évaluer les technologies développées et les intégrer dans une nouvelle version de la plateforme INFOM@GIC[2] pour en dériver trois chaînes de traitement applicatives dédiées à la veille d'opinion, l'intelligence du consommateur et du citoyen, et à la fidélisation et attrition de clientèle pour les end-users du projet : OpinionWay, EDF et Meetic,
– évaluer les usages des utilisateurs finaux DoXa et développer de nouveaux services pour outiller les acteurs des marchés visés.

Les chaînes de traitement développées pour les end-users aideront à l'observation dynamique, quantitative et qualitative, du positionnement des consommateurs, clients, utilisateurs, des rapports qu'ils entretiennent avec les univers à propos desquels ils s'expriment ainsi que les tendances ou évolutions à l'œuvre dans ces univers. Elles aideront à améliorer les processus d'analyse décisionnelle et de fouille (OLAP [3], segmentation, calcul de score, etc.) en intégrant dans ces processus des connaissances enrichies.

Les corpus du projet sont constitués d'une part, de corpus de blogs qui portent sur les thèmes des jeux vidéos, de la crise économique et financière et de la maîtrise de l'énergie, d'autre part, de corpus de remontées clients des services consommateurs.

2 Les modèles d'analyse de sentiment et d'opinion

Le vocable de modèle d'analyse de sentiment et d'opinion (ASO) regroupe des approches très variées, de manière générale soit orientées vers la découverte des expressions d'opinions basées sur des considérations plus ou moins rationelles, des jugements ou des appréciations, soit orientées vers la modélisation des sentiments et des émotions qu'une personne peut entretenir à propos d'un objet particulier. Les modèles varie beaucoup dans la nature et le nombre de dimensions qu'ils mettent en œuvre, ainsi que dans la granularité de leur représentation sémantique.

D'après Esuli et Sebastiani (2006), la fouille d'opinion consiste à la fois à rechercher les expressions des sentiments et opinions dans les documents et à acquérir de nouvelles méthodes pour effectuer automatiquement de telles analyses. Ils mentionnent en particulier trois aspects importants :

1. développer des ressources langagières pour la fouille d'opinion, par ex. des lexiques ou des thésaurus de termes subjectifs ;
2. classer des textes en fonction des opinions qu'ils expriment ;
3. extraire des textes les expressions d'opinion en prenant en compte la relation qui lie les expressions d'opinion (les mots explicitant l'opinion), la source (l'auteur de l'expression d'opinion) et la cible (l'objet de l'opinion exprimée), Kim et Hovy (2006).

Pour organiser notre état de l'art, nous avons d'abord fait la liste des attributs que nous percevions comme saillants à la lecture des descriptions des modèles et les avons rangés selon l'or-

[2] Infom@gic est un projet du pôle Cap Digital, consacré à l'analyse de l'information multi-modale ; le projet a démarré en 2006 et s'est achevé en 2009.

[3] *On-line Analytical Processing*, analyse interactive de grandes quantités de données multidimensionnelles.

dre suivant d'importance relative décroissante par rapport à l'expression d'opinion : *la polarité*, *l'intensité*, *la cible*, *l'informativité* (le caractère plus ou moins factuel de l'expression d'opinion), *l'engagement* et *la source*. Nous avons ensuite classé les modèles en fonction du nombre d'attributs qui leur était associé[4]. Par exemple, nous avons considéré qu'un modèle qui disposerait uniquement de l'attribut *polarité* serait plus générique, qu'un modèle qui prendrait en compte une dimension supplémentaire comme la *cible*.

Les modèles les plus génériques, ceux de Quirk et al. (1985), Kamps et al. (2004) et Berthard et al. (2004) n'ont pas proposé d'attribut en rapport directe avec notre contexte applicatif, mais ont adressé de manière générale le problème de l'opinion et des sentiments dans le language. Quirk et al. (1985) a introduit la notion d'état privé (*private state*) qui regroupe toutes les expressions de subjectivité comme les émotions, les opinions, les attitudes, les valuations etc. Cette notion est aussi présente dans les modèles de Wiebe et al. (2005) et Pang et Lee (2008). Puis viennent au niveau suivant de la hierarchie, les modèles de Dave et al. (2003), Turney (2002), Har, Somasundaran et al., Kim et Hovy (2006) et Stoyanov et al. (2007). Ensuite nous trouvons au niveau suivant les modèles de Mullen et Collier (2004), Stoyanov et al. (2007) et Yu et Hatzivassiloglou (2003) qui prennent en compte la source et la cible d'une expression d'opinion. Le travail de Yannik-Mathieu (1991) est caractérisé par un classement des verbes de sentiments.

Le modèle de Martin et White (2005) s'intéresse aux aspects évaluatifs. Les auteurs ont mentioné 3 types d'évaluation caractérisés en fonction de leurs attributs : *attitude*, *engagement* et *gradation* . *Attitude* se rapporte aux valeurs de jugement issues d'un ou plusieurs sources et peut-être associée à une réponse émotionnelle. Ses trois sous types sont : *jugement*, *affect* et *appreciation*. L'*engagement* explicite la position, la plus ou moins grande implication de la source par rapport à l'expression de son opinion. C'est une des caractéristiques principales de la subjectivité. La notion de *gradation* peut se décliner en deux dimensions *force* et *focus* (Martin et White (2005)), celles-ci se retrouvent regroupées dans notre terminologie sous la notion de *l'intensité*. Les modèles de Choi et al. (2005) et Riloff et al. (2003) se placent dans notre hiérarchie au même niveau que ceux de Riloff et al. (2006), Turney et Littman (2003) et Yannik-Mathieu (1991). D'après nous, les modèles les plus riches sont ceux de Pang et Lee (2008) and Wiebe et al. (2005) qui regroupent ensemble tous les attributs de notre système de classement.

De cette présentation de l'état de l'art, il ressort une distinction entre deux courants, d'une part l'analyse de sentiment, plus liée à l'étude des émotions et d'autre part l'analyse d'opinion qui est plus en relation avec des aspects liés au raisonnement logique ou à l'appréciation. En outre, nous retenons que les modèles de fouille d'opinion s'organisent autour de trois composantes essentielles, la source, la cible et l'expression d'opinion. Nous allons maintenant faire un bref rappel sur les dernières campagnes d'évaluation en terme de fouille d'opinion.

3 Les campagnes d'évaluation

Le domaine de la fouille de sentiment et d'opinion n'échappe pas à l'engouement actuel en traitement automatique des langues pour le paradigme d'évaluation Paroubek (2009) comme les exemples suivants l'attestent.

[4]Mlle Thuy Linh Ngo a contribué à l'élaboration de cette hierarchie lors de son stage de fin de master au LIMSI pendant l'été 2009.

De 2006 à 2008, TREC (Text Retrieval Conference) a proposé le "*Blog Track*"[5]. La tâche consistait d'abord à distinguer si un blog était de nature objective (pas d'opinion exprimée) ou subjective et ensuite à séparer les blogs en fonction de leur polarité, en les classant par ordre décroissant de positivité. Les performances ont été mesurées en terme de "*Mean Average Precision (MAP)*", avec des valeurs comprises entre 0,17 et 0,45.

Depuis 2006, la campagne NTCIR-MOAT (*Multilingual Opinion Analysis Task*) propose d'étiqueter des articles de presse. Chaque phrase doit être étiquetée en fonction de sa plus ou moins grande subjectivité et de sa pertinence par rapport à un thème prédéfini. Les participants doivent identifier la source et la cible d'une expression d'opinion au niveau sous-phrastique (clause) ainsi que sa polarité codée sur une échelle à trois valeurs : positive, négative et neutre. Dans cette campagne, l'identification de la subjectivité et de la pertinence a été mieux réussie (F-scores allant de 0.41 à 0.92) que l'identification de la source, la cible et la polarité d'une expression d'opinion (F-scores entre 0 et 0.75).

En 2007, la campagne SemEval (Semantic Evaluations) a proposé la tâche *Affective Text* pour explorer les liens entre les expressions d'émotions et les lexiques sémantiques. Les participants devaient annoter les gros titres des journaux selon deux dimensions, l'émotion et la polarité. Les émotions les mieux reconnues ont été : la tristesse et la peur avec des F-scores maximaux respectifs de of 0.30 et 0.20. La colère, la joie et la surprise ont obtenu un F-score maximum de 0.15, tandis que le dégoût a été l'émotion la moins bien reconnue avec un F-score nul.

Finalement en 2008, la campagne TAC (Text Analysis Conference) avait une tâche de question-réponse portant sur des questions d'opinion. Il y avait deux types de questions, un groupe portait sur la source de l'opinion et le second portait sur l'opinion elle-même. Les F-scores obtenus pour ces deux types de question se situaient dans l'intervalle 0.01-0.17.

En France deux campagnes d'évaluation ont eu lieu, respectivement en 2007 Grouin et al. (2007) et 2009 [6], dans le cadre de DEFT, le DÉfi Fouille de Texte.

La première campagne portait sur l'affectation d'une valeur de polarité sur des critiques de livres, de films ou de jeux vidéo Jean-Baptiste Berthelin (2008). La seconde campagne proposait aux participants de déterminer si un texte était globalement de nature subjective ou objective, puis d'identifier les portions subjectives, ceci dans un corpus d'articles de journaux.

Dans la campagne de 2007 les meilleurs résultats ont été obtenus avec le corpus de critiques de jeux-vidéo, avec des valeurs de F-score allant de 0.46 à 0.78, tandis que le corpus de relecture d'articles scientifique est celui qui a posé le plus de problèmes aux participants, avec des F-scores compris entre 0.40 et 0.57.

En 2009, les participants ont obtenu un succès légèrement meilleur sur la tâche d'analyse globale (F-scores allant de 0.66 à 0.92 sur le corpus français) qu'avec la tâche d'identification des portions subjectives (F-score allant de 0.65 à 0.91).

A notre avis, il y a deux points à retenir de ce bref survol des activités passées d'évaluation concernant l'analyse de sentiment et d'opinion :

1. la diversité des tâches proposées dans les différente campagnes est représentative de la diversité des approches existantes, il n'existe donc pas de protocole « standard »,

[5] http ://ir.dcs.gla.ac.uk/wiki/TREC-BLOG/
[6] http ://deft09.limsi.fr

2. pour peu que l'on se donne un contexte d'application suffisament bien défini, l'analyse de sentiment et d'opinion automatique se prète bien au jeu de l'évaluation technologique en mode quantitatif boîte-noire.

4 Présentation du modèle Opinions et Sentiments de DOXA

L'un des défis du projet DoXa porte sur la modélisation des opinions et sentiments permettant d'aller au-delà de la simple opposition Positif/Négatif. Aujourd'hui, les approches académiques de même que les solutions du marché permettant d'analyser automatiquement les opinions et sentiments s'appuient sur des méthodes de catégorisation qui structurent les textes analysés, en fonction du jugement favorable ou défavorable exprimé. Ce résultat est un premier résultat qui est certes intéressant mais qui est insuffisant au regard des besoins métier, dans la mesure où il manque de finesse et de précision. Les analystes ont en effet besoin de caractériser plus précisément et plus finement les opinions et sentiments qui relèvent du positif et du négatif. Que peut-on déduire en effet d'une analyse qui indique que sur un corpus de 10 000 interactions clients, 8000 interactions relèvent d'un discours négatif ? Quelle décision peut-on prendre à partir d'un tel résultat ? Cette caractérisation effectuée, l'analyse humaine doit prendre le relais pour l'exploiter, car il faut l'affiner en montrant comment se structurent les discours positif et négatif.

Dans le cadre du projet, nous avons défini un modèle d'analyse des opinions et sentiments[7], qui s'appuie sur la fusion d'un modèle sémantique - fondé sur un ensemble de catégories sémantiques d'opinions et de sentiments et sur un ensemble de traits - et d'un modèle numérico-symbolique - fondé sur calculs à base de logique floue[8]. Nous avons en parallèle, suite à l'expression de besoins des end-users, défini un mode de représentation des textes qui prend en compte trois niveaux d'analyse du texte :

– Le niveau micro qui correspond au niveau phrase ou segment de phrase,
– Le niveau meso qui correspond au niveau du paragraphe,
– Le niveau macro qui correspond au niveau du texte.

Ces niveaux d'analyse utilisés pour l'annotation en thèmes et en opinions et sentiments permettent de mettre en œuvre un parcours d'analyse allant de la vision détaillée (micro) à la vision synthétique (macro).

Le modèle d'analyse sémantique des opinions et sentiments est appliqué au niveau micro ; l'analyse est réalisée par des méthodes symboliques à base de ressources lexicales et de grammaires locales. Le modèle numérico-symbolique, prend en entrée les annotations en catégories sémantiques et en traits du niveau micro, il est appliqué aux niveaux meso et macro, en vue d'effectuer une synthèse d'informations.

La liste des catégories sémantiques a été définie en s'appuyant d'une part, sur les travaux de Martin and White, pour les valeurs intellectives et sur les travaux de Yannick Yvette Mathieu, pour les valeurs affectives, d'autre part, sur l'ensemble des corpus du projet et sur les attentes exprimées par les end-users. Les travaux d'implémentation du modèle par les linguistes de

[7]Le modèle Opinions et Sentiments du projet a initialement été défini par Thales, pour la partie sémantique et le LIP6, pour la partie numérico-symbolique. Dans sa version finale, il est le résultat dćun étroite et fructueuse collaboration entre Thales, le LIP6, Arisem, IGM et le LIMSI.

[8]Un article présentant de manière plus détaillée le modèle élaboré et son implémentation est en cours de rédaction ; il sera publié dans le courant de l'année 2010.

même que les travaux d'annotation manuelle en vue de l'évaluation technologique ont permis d'optimiser la liste définie initialement.

Au niveau micro, tous les segments porteurs d'opinions ou de sentiments sont annotés via le jeu de catégories sémantiques. Aux niveaux meso et macro, on effectue une synthèse des catégories sémantiques en se donnant une limite de 5 catégories sémantiques, ceci en vue de produire des annotations synthétiques. La synthèse est produite compte tenu des relations de proximité sémantique des catégories et de la fréquence d'apparition de ces dernières.

Pour compléter l'analyse en **catégories sémantiques**, un ensemble de **traits** est utilisé aux trois niveaux d'analyse :

Au niveau micro : la polarité, le neutre lesquels sont déduits automatiquement à partir des catégories sémantiques définies ; l'intensité, la négation et l'ambiguïté.

La **polarité** est l'orientation positive ou négative de l'opinion exprimée dans le texte. Un texte objectif ou sans opinion est dit **neutre**.

L'**intensité** est la force de l'orientation positive ou négative. Ce trait prend 10 valeurs numériques allant de 1 à 10.

La **négation** permet de traiter les catégories sémantiques sans antonyme telle que la colère. Ainsi un segment du type « Je ne suis pas en colère »sera annoté en Négation de COLERE.

L'**ambiguïté** permet de traiter les segments ambigus pour lesquels le système d'analyse n'est pas capable de prendre de décision, en termes de catégories.

Aux niveaux meso et macro : on fusionne les traits polarité et intensité selon 5 valeurs et l'on conserve le trait **neutre**.

Un complément du modèle Opinions et Sentiments :
Les corpus étudiés dans le cadre du projet contenant des textes de type réclamations, nous avons été amenés à ajouter à l'axe des opinions et sentiments, deux axes correspondant à des actes de langage Colletta (1998) : l'axe de la demande et l'axe de la recommandation ou de la suggestion. Ainsi si l'énonciateur exprime une opinion ou un sentiment, le segment évaluatif sera annotée selon l'ensemble des catégories sémantiques ayant trait aux jugements, sentiments et émotions. Si l'énonciateur exprime une demande, le segment porteur de la demande sera annoté selon la catégories Demande_Requete ; si l'énonciateur exprime une suggestion, le segment porteur de la suggestion sera annotée selon la catégorie Recommandation_Suggestion. Aux niveaux meso et macro, les trois axes ne sont évidemment pas exclusifs.

L'analyse thématique est comme on a l'a indiqué précédemment appliquée aux trois niveaux du texte. Le thème est défini comme l'entité concrète ou abstraite qui est la cible de l'opinion, ce sur quoi porte l'opinion ou le sentiment. Le segment thématique prend ses valeurs dans l'ontologie des thèmes (ontologie métier). C'est l'ontologie qui permet de faire la synthèse des thèmes du niveau micro au niveau macro.

5 Le corpus de référence

Les corpus sont fournis par les partenaires industriels du projet DOXA. Ces corpus recouvrent trois domaines, les jeux vidéos, la crise économique, et les réclamations-clients. Les deux premiers corpus ont été constitués à partir de données d'Internet, alors que le troisième est constitué de données propres aux partenaires. A chacun de ces domaines est associée une ontologie-métier regroupant les termes propres au domaine. Ces ontologies sont construites par les partenaires du projet.

Chaque corpus doit être annoté par des annotateurs humains sur la base du modèle d'annotation défini dans le projet DOXA. Dans cet article, nous nous intéressons plus spécifiquement au corpus des jeux vidéos, qui est le premier à être annoté. Ce corpus a été collecté à partir de 8 sites spécifiques. Il contient un mélange de critiques et journaux provenant de professionnels du domaine, et de blogs d'internautes ou de messages publiés sur des forums dédiés aux jeux vidéo. Il y a environ 8,000 documents d'une taille moyenne de 4500 mots, qui pour les besoins du projet DOXA ont été découpés en paragraphes de taille fixe et codés en UTF8.

Exemple de paragraphe : *<Parag id="d1009.1"> Test Arthur Et Les Minimoys VidéoTest En général, nous redoutons les adaptations en jeu d'un livre ou d'un film. Nous avons tellement été déçus par des transpositions ne cherchant qu'à tirer profit d'une licence au détriment de la qualité,que nous avions toutes les raisons d'appréhender l'arrivée d'Arthur et les Minimoys, adaptation des romans et du film de Luc Besson du même nom. Le studio français Étranges Libellules s'est lancé dans l'aventure pour nous proposer, au final, un titre véritablement abouti et soigné. </Parag>*

Le corpus de référence doit servir en partie à l'entraînement des logiciels développés dans le projet, et pour une autre partie, à leur évaluation. Nous avons vu dans la section 4 que le modèle d'opinions et sentiments s'applique à trois niveaux de texte, les niveaux macro (document), méso (paragraphe) et micro (phrase). L'annotation humaine est un travail assez lourd et nous avons considéré qu'annoter les trois niveaux pénaliserait la taille du corpus de référence. Nous avons donc choisi d'annoter les deux premiers niveaux, macro et méso, en laissant de côté le niveau micro, car les niveaux synthétiques sont les plus intéressants dans un cadre d'exploitation des résultats. Chaque document est donc annoté manuellement à deux niveaux, le niveau macro (le document) et le niveau méso (chaque paragraphe).

Les attributs à annoter sont ceux définis dans la section 4 : la *catégorie sémantique* et les traits *polarité*, *intensité* et *thème*. Les attributs *thème* et *catégorie sémantique* prennent leurs valeurs respectivement dans les ontologies-métier et dans l'ensemble des catégories sémantiques. L'attribut *polarité* prend une valeur dans l'ensemble *positif*, *négatif*, *mixte*, *neutre*, et l'attribut *intensité* prend l'une des valeurs *fort* ou *moyen* lorsque la polarité est positive ou négative, et aucune valeur sinon. Nous avons préféré garder séparés la *polarité* et l'*intensité* pour l'annotation manuelle, la fusion étant effectuée à posteriori pour retrouver les 5 valeurs définies dans le modèle. Les annotations sont justifiées par les parties de texte qui les représentent le mieux. Ainsi, à la valeur de l'attribut *thème* sera associé l'extrait de texte parlant de ce thème, et à la valeur de l'attribut *catégorie sémantique* sera associée l'expression illustrant cette valeur dans le texte. Un attribut complémentaire a été ajouté aux attributs du modèle : c'est l'attribut *justification* dont la valeur est l'extrait de texte qui illustre le mieux l'ensemble de l'annotation. Ces différentes justifications ont un double but : d'abord un but d'autovérification, car l'annotateur doit pouvoir expliciter ses choix. Le second but est de recueil de données, car ainsi, à chaque valeur d'opinion sera associé un ensemble de segments de textes la représentant.

Exemple d'annotation :

Parag id="d1009.1"	
Polarité :	*positif*
Intensité :	*fort*
Catégorie sémantique :	*Interet_Valorisation_Appreciation Texte : véritablement abouti*
Thème :	*VideoGame Texte : Arthur Et Les Minimoys*
Justification :	*Le studio français Étranges Libellules s'est lancé dans l'aventure pour nous proposer, au final, un titre véritablement abouti et soigné.*

6 L'interface d'annotation

Nous avons utilisée Knowtator Ogren (2006) [9], un plugin de Protégé qui permet de disposer d'une interface graphique d'annotation couplée à un éditeur d'ontologie. Cette interface

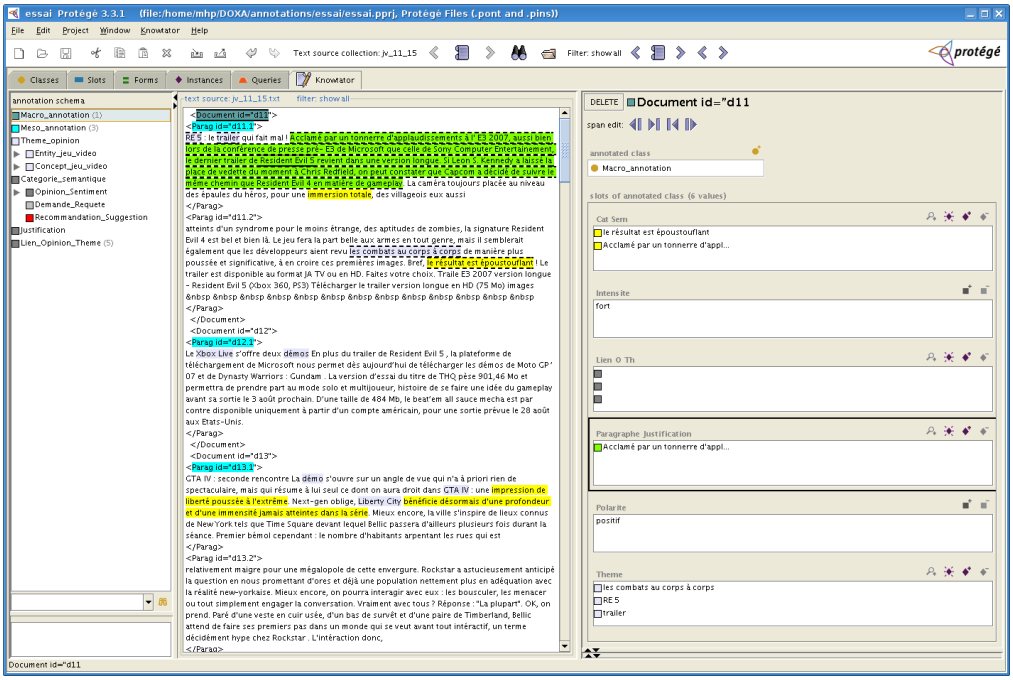


FIG. 1 – *Interface d'annotation avec Knowtator.*

a été choisie car elle permet d'annoter des éléments de texte en se servant des valeurs d'une ontologie définie à l'avance. Cette ontologie se présente comme une hiérarchie de concepts. Ainsi, nous avons pu intégrer à cette interface l'ontologie-métier des jeux vidéo et la hiérarchie des catégories sémantiques d'opinion et de sentiments définies par les partenaires du projet. Toutes les valeurs des attributs de l'annotation sont donc définies à l'avance, et les annotateurs affectent ces valeurs à des parties de texte.

Le logiciel Knowtator étant dédié à l'annotation d'un texte au niveau micro, il nous faut adapter son fonctionnement à l'annotation des niveaux macro et méso. En effet, la démarche générale à suivre dans Knowtator est d'abord de sélectionner une partie de texte, puis de lui affecter des valeurs d'attributs d'annotation. Pour les niveaux macro et méso, il faudrait donc sélectionner le document, ou le paragraphe, dans son intégralité, pour lui affecter ensuite les annotations. Les manipulations d'annotation seraient assez lourdes, et nous aurions les textes in extenso, mais pas leurs références en termes de numéro de document ou de paragraphe. La solution que nous avons adoptée est d'importer le texte à annoter sous le format XML, et de

[9] http ://knowtator.sourceforge.net/

sélectionner pour le niveau méso l'étiquette XML signalant le paragraphe et son numéro, et pour le niveau macro l'étiquette XML signalant le document et son numéro. Ainsi, les valeurs des attributs d'annotation seront reliés aux étiquettes XML des débuts de document ou de paragraphe, et non pas au texte entier. Les segments de texte des justifications seront en revanche sélectionnés dans leur intégralité et reliés aux niveaux macro ou méso correspondants.

L'interface d'annotation de Knowtator-Protégé comporte une partie à gauche contenant le schéma d'annotation intégrant les ontologies de domaine et de catégories sémantiques, une partie au milieu contenant le texte à annoter, et une partie à droite permettant de visualiser les annotations (voir figure 1). Le texte à annoter est sélectionné dans la partie du milieu, les deux autres parties permettant de créer et d'éditer les annotations.

7 Accords inter-annotateurs et règles d'annotation

7.1 Accords

Nous avons fait annoter le même ensemble de textes par deux annotateurs différents afin de détecter leurs accords et désaccords et d'établir des règles d'annotation permettant de limiter les divergences entre annotateurs. Le tableau 1 montre des valeurs d'accord pour différents ensembles de documents. Des réunions nous ont permis de faire le point périodiquement et

%=	0.62			0.72			0.78			0.87			0.82		
# docs	56			50			63			48			34		
	A1	A2	=	A1	A2	=	A1	A2	=	A1	A2	=	A1	A2	=
+	19	19	14	21	26	17	22	28	19	20	21	18	8	7	6
-	19	17	12	13	13	10	22	17	17	15	14	13	19	20	17
+/-	5	1	0	0	0	0	1	0	0	0	0	0	1	1	0
~	13	19	9	16	11	9	18	18	13	13	13	11	6	6	5

TAB. 1 – *Exemples de valeurs d'accord pour différents ensembles de documents, avec : pourcentage d'accord (%=) ; nombre de documents (# docs) ; A1,A2 (annotateurs) ; nombre d'accords interannotateurs (=) ; polarité positive (+) ; polarité négative (-) ; polarité mixte (+/-) ; polarité neutre (~).*

d'établir un ensemble de règles et de recommandations dont un extrait est donné en section 7.2. Par ailleurs, des listes d'expressions textuelles traduisant les différentes valeurs de catégories sémantiques et de thèmes ont été élaborées à partir des textes associés par les annotateurs aux valeurs donnés aux attributs catégorie sémantique et thème pour chaque document. Ces listes servent de référence aux annotateurs.

7.2 Règles d'annotation

Les règles suivantes donnent les caractéristiques que le texte doit présenter pour que les attributs de polarité et d'intensité de l'opinion émise dans le texte, aient les valeurs indiquées. Les contraintes sur les catégories sémantiques, étant donné ces valeurs d'attributs, sont aussi exprimées.

Polarité

- positif
  - caractéristiques du texte : le texte exprime une opinion, les jugements sont en majorité favorables, les sentiments exprimés sont en majorité positifs.
  - valeurs d'intensité :
    - fort
      - caractéristiques du texte : les expressions de l'opinion sont très positives ET les expressions négatives sont absentes OU faiblement présentes ET très modérées.
      - valeurs des catégories sémantiques associées : pas de catégorie sémantique négative, uniquement des catégories sémantiques positives.
    - moyen
      - caractéristiques du texte : 1) les expressions de l'opinion sont moyennement positives ET les expressions négatives sont absentes OU faiblement présentes et très modérées OU 2) les expressions sont positives à très positives ET Les expressions négatives sont présentes mais très modérées OU faiblement présentes et modérées.
      - valeurs des catégories sémantiques associées : des catégories sémantiques positives, et éventuellement une catégorie sémantique négative.
- negatif *symétrique du positif*
- mixte
  - caractéristiques du texte : le texte exprime une (des) opinion(s), et 1) les jugements sont de tonalité moyenne, ni franchement positifs ni franchement négatifs, et les sentiments exprimés sont très mesurés, OU 2) les jugements ainsi que les sentiments exprimés sont contrastés en positif/négatif. Si l'opinion est contrastée, les marqueurs d'opposition-concession (mais, si, malgré, etc.) sont très présents.
  - valeur additionnelle de polarité : une deuxième valeur de polarité peut être donnée : positif si la tendance est légèrement favorable, négatif si la tendance est légèrement défavorable.
  - valeurs d'intensité : aucune valeur
  - valeurs de catégories sémantiques associées : au moins une catégorie sémantique positive ET une catégorie sémantique négative.
- neutre
  - caractéristiques du texte : le texte n'exprime pas d'opinion.
  - valeurs d'intensité : aucune valeur
  - valeurs de catégories sémantiques associées : les seules catégories sémantiques possibles sont Demande_Requete et Recommandation_Suggestion. Si elles ne sont pas pertinentes : aucune valeur de catégorie sémantique.

8 Conclusion

Nous venons de présenter la construction d'un corpus de référence pour l'évaluation de systèmes de fouille d'opinion dans le cadre du projet industriel DOXA. Après avoir fait le tour des différentes approches de l'analyse de sentiment et d'opinion et effectué un retour sur les principales campagnes d'évaluation, nous avons décrit les outils et principes d'annotation du corpus de référence issu du WEB concernant les jeux vidéo. Il est à noter que le seul paramètre dépendant du domaine d'application dans le methodologie décrite et les outils utilisés est l'ontologie métier, faisant de l'ensemble un système portable de construction de données de référence (*gold standard*) pour l'évaluation de la fouille d'opinion.

Références

Berthard, S., H. Yu, A. Thornton, V. Hativassiloglou, et D. Jurafsky (2004). Automatic extraction of opinion propositions and their holders. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text*.

Choi, Y., C. Cardie, E. Riloff, et S. Patwardhan (2005). Identifying sources of opinions with conditional random fieldsand extraction patterns. In *Proceedings of HLT/EMNLP*.

Colletta, J.-M. (1998). A propos de la modalisation en franĄçais oral. In G. Ruffino (Ed.), *Atti del XXI Congresso Internazionale di Linguistica et Filologia Romanza*, TĄübingen, pp. 65–80. Universita du Palermo : Niemeyer Verlag.

Dave, K., S. Lawrence, et D. M. Pennock (2003). Mining the peanut gallery :opinion extraction and semanctic classification of product reviews. In *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary.

Esuli, A. et F. Sebastiani (2006). Sentiwordnet : A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, Genova, Italy, pp. 417–422.

Grouin, C., J.-B. Berthelin, S. E. Ayari, T. Heitz, M. Hurault-Plantet, M. Jardino, Z. Khalis, et M. Lastes (2007). Présentation de deft'07 (defi fouille de textes). In *Actes de l'atelier de clôture du 3ème DEfi Fouille de Textes*, Grenoble, pp. 1–8. Association Française d'Intelligence Artificielle.

Jean-Baptiste Berthelin, Cyril Grouin, M. H.-P. P. P. (2008). Human judgement as a parameter in evaluation campaigns. In *Proceedings of the Coling Workshop on Human Judgements in Computational Linguistics (HJCL 2008)*, Manchester.

Kamps, J., M. Marx, R. J. Mokken, et M. de Rijke (2004). Using wordnet to measure semantic orientation of adjectives. In *Proceedings of LREC*, Volume IV, pp. 174–181.

Kim, S.-M. et E. Hovy (2006). Identifying and analyzing judgment opinions. In *Proceedings of the Joint HLT / NACACL Conference*.

Martin, J. R. et P. R. R. White (2005). *The Language of Evaluation :Appraisal in English* (illustrated ed.). Palgrave Macmillan.

Mullen, T. et N. Collier (2004). Sentiment analysis using support vector machines diverse information sources. In *Proceedings EMNLP-04*.

Ogren, P. (2006). Knowtator : A protégé plug-in for annotated corpus construction. In A. for Computational Linguistics (Ed.), *Proceedings of the Conference of the North American Chapter of the ACL on Human Language Technology : companion volume : demonstrations*, New-York, pp. 273–275.

Pang, B. et L. Lee (2008). *Opinion mining and sentiment analysis*, Volume 2. now, the essence of knowledge.

Paroubek, P. (2009). Evaluation : a paradigm that produces high quality language resources. In *Proceedings of the FlaReNet Forum - The European Language Resource and Technology Forum : Shaping the future of the Multilingual Digital Europe*, Vienna, pp. 64–66. Istituto Di Linguistica Computazionale del CNR.

Quirk, R., G. Leech, et J. Startvik (1985). *A Comprehensive Grammar of the English Language*. NewYork :Longman.

Riloff, E., S. Patwardhan, et J. Wiebe (2006). Feature subsumption for opinion analysis. In *In : Proceedings of EMNLP*.

Riloff, E., J. Wiebe, et T. Wilson (2003). Learning subjective noun using extraction pattern bootstrapping. In *Proceedings of the 7th Conference on Natural Language Learning*, Edmonton, Canada, pp. 25–32.

Somasundaran, S., J. Ruppenhofer, et J. Wiebe. Discourse level opinion relations :an annotation study. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pp. 129–137. Association for Computational Linguistics.

Stoyanov, V., C. Cardie, D. Littman, et J. Wiebe (2007). Evaluating an opinion annotation scheme using a nex multi-perspective question and answer corpus. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*. AAAI Technical Report SS.

Turney, P. (2002). Thumbs up or thumbs down ?semantic orientation applied to unsupervised classification of reviews. In *Proceedings of 40th Meeting of the Association for Computational Linguistics*, pp. 417–424.

Turney, P. et M. Littman (2003). Measuring praise and criticism :inference of semantic orientation from association. pp. 315–346. ACM Transactions on Information Systems.

Wiebe, J., T. Wilson, et C. Cardie (2005). Annotating expressions of opinions and emotions in language. Netherlands : Kluwer Academic Publishers.

Yannik-Mathieu, Y. (1991). Sciences du langage. In *Les verbes de sentiment – De l'analyse linguistique au traitement automatique*. CNRS Editions.

Yu, H. et V. Hatzivassiloglou (2003). Towards answering opinion questions :separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP*, Sapporo, Japan, pp. 129–136.

Summary

We present the building of a reference corpus for opinion mining evaluation in the industrial context of the DOXA project. First we describe the DOXA project and make a short presentation of the state-of-the-art about sentiment and opinion analysis before looking back at recent evaluation campaigns. Then we describe the corpus about video games that we have collected from the WEB and its annotation scheme. Before concluding, we present the annotation interface based on knowtator.

Regroupement sémantique de définitions en espagnol

Gerardo Sierra[1], Juan-Manuel Torres-Moreno[2,3], Alejandro Molina[2]

[1]Universidad Nacional Autónoma de México, Ciudad Universitaria, México D.F.
gsierram@iingen.unam.mx http://www.iling.unam.mx

[2]Laboratoire Informatique d'Avignon, BP1228, 84911 Avignon Cédex 09, France
juan-manuel.torres@univ-avignon.fr alejandro.molina@etd.univ-avignon.fr
http://lia.univ-avignon.fr/

[3]École Polytechnique de Montréal, CP 6079 Succ. Centre-ville, Montréal (Québec) Canada

Résumé. Cet article s'intéresse à la description et l'évaluation d'une nouvelle méthode d'apprentissage non supervisé pour réunir des définitions en espagnol selon leur signification. Nous utilisons comme mesure de regroupement l'énergie textuelle et nous étudions une adaptation de la précision et le rappel afin d'évaluer notre méthode.

1 Introduction

De nos jours, l'utilisation de l'Internet pour la recherche de définitions est de plus en plus importante. Wikipédia et Medline sont devenu les sites les plus consultés de la Web[1]. Or, il existe un énorme nombre de définitions qui sont parfois inaccessibles aux utilisateurs. Celles-ci peuvent se trouver dans des sites non encyclopédiques ou dans de documents divers. Dans cette perspective nous avons développé le moteur de recherche *Describe*[2], qui permet de trouver des définitions en espagnol (Sierra et al., 2009). Une caractéristique de ce moteur est qu'il regroupe les résultats des recherches (définitions liées à un terme). Cet article présente la méthodologie de regroupement et l'évaluation des résultats. Ceux-ci sont encourageants du point de vue qualitatif. Par contre, l'évaluation quantitative pose des contraintes car il est compliqué d'évaluer la sémantique.

Cet article est organisé comme suit : dans la section 2 nous introduisons les contextes définitoires (CD), dans la section 3 nous présentons des stratégies de regroupement des définitions. Le corpus utilisé dans nos expériences est présenté en section 4. Des évaluations avec des analyses quantitative et qualitative sont présentées au chapitre 5 avant de conclure et de donner quelques perspectives.

2 Contextes définitoires

Un contexte définitoire (CD) est un fragment textuel qui définit un terme (Alarcón et al., 2009). Par exemple: *Le navire, peut être originellement considéré comme un flotteur qui*

[1] http://www.alexa.com/topsites
[2] http://www.describe.com.mx

essaie de rester dans une position verticale face à des perturbations extérieures. Le fragment précédent est un CD car il possède le terme (T) : *navire* et sa définition respective (D) : *un flotteur qui essaie …* Dans cet exemple, le terme et sa définition se rattachent grâce à la structure syntaxique *peut être+considéré+comme*. Ce type de structure est nommée patron définitoire (PD). Mis à part le terme et la définition, un CD possède d'autres éléments ayant déjà été définis dans d'autres travaux (Sierra et al., 2004). Dans cet article il suffit de connaître les deux éléments auparavant décrits.

Les PDs peuvent être des structures syntaxiques bien spécifiques, comme la séquence : *peut+se+définir+comme*[3] (Rodríguez, 2004). Chaque PD peut s'associer à un type de définition établie selon la relation entre T et D. L'extraction de CDs moyennant des PDs en français a été étudiée par Malaisé (2005) et Rebeyrolle et Tanguy (2000). En espagnol, il y a quatre types de définitions basées sur le modèle aristotélique : l'analytique, l'extensionnel, le fonctionnel et le synonymique. Il existe des études approfondies des trois premiers (Aguilar et Sierra, 2009), pour cette raison nous laissons de côté le modèle synonymique.

Les définitions analytiques présentent un genus qui exprime la catégorie la plus générale à laquelle le terme appartient et une différentia qui permet de le distinguer d'autres éléments de la même catégorie. Quelques patrons verbaux liés à ce type de définition sont : *être+un, définir+comme, comprendre+comme*. Les définitions extensionnelles énumèrent les parties qui conforment le terme définit. Quelques verbes liés aux définitions extensionnelles sont : *contenir, comprendre,* et *inclure*. Les définitions fonctionnelles ne présentent pas un genus, mais une différentia qui exprime l'usage particulier du terme est introduite. Quelques patrons liés à ce type de définitions sont : *fonctionner, permettre* et *sert+à*.

3 Regroupement de définitions

3.1 Regroupement hiérarchique

On introduira d'abord quelques définitions nécessaires à la compréhension du reste de la section. Un document est une chaîne fini de longueur arbitraire de symboles graphiques nommés entités lexicales (EL). Une EL est soit un symbole soit l'union de plusieurs symboles. Par exemple, le mot *pomme* ou un groupe de mots comme *République Française.* De même, une EL peut être un symbole inintelligible comme *Viv* ou *B4.* Une collection est un ensemble de documents et un dictionnaire est la liste de ELs uniques apparaissant dans une collection.

La représentation des définitions est basée sur le modèle vectoriel de Salton (1971). Les définitions sont des vecteurs avec le même nombre de dimensions que d'ELs dans le dictionnaire. Néanmoins, étant donné que les définitions sont généralement courtes, nous avons utilisé l'approche présence/absence de termes : 1 si l'EL apparaît et 0 autrement. La collection des définitions peut être donc représentée par une matrice binaire document-EL. Evidement, cette représentation empêche l'utilisation de mesures de similarité basées sur des pondérations réelles (cosinus par exemple).

Le regroupement hiérarchique offre un grand avantage : le nombre de classes ne doit pas être spécifié. Dans la version la plus simple (HAC) l'entrée est un ensemble d'objets et la sortie un dendrogramme, c'est-à-dire un arbre hiérarchique qui regroupe tous les objets.

[3] Traduction en français de *se+puede+definir+como*.

Dans chaque itération, une fonction calcule la distance entre chaque paire de groupes afin de déterminer les deux groupes à fusionner. Le critère pour calculer la distance entre chaque groupe est généralement une variante des méthodes *linkage*. Dans la méthode du *complete linkage* (Sorensen, 1948), la distance entre les groupes est représentée par la plus grande distance entre un objet du premier groupe et un objet du deuxième. Cette méthode a l'avantage de créer de petits groupes, cohésifs et bien délimités. C'est la raison principale pour laquelle nous l'avons utilisée. La distance entre le groupe D_i et le groupe D_j, où $d_i \in D_i$ et $d_j \in D_j$, est définie par :

$$Dist(D_i, D_j) = \max_{d_i \in D_i, d_j \in D_j} dist(d_i, d_j) \quad (1)$$

où *dist* est une fonction dont l'ensemble de départ sont les objets, à l'opposé de *Dist* dont l'ensemble de départ sont les groupes.

3.2 La valeur de seuil par distance

Pour obtenir un ensemble de groupes, nous utilisons un seuil, afin de stopper les fusions du dendrogramme. De cette façon, on obtient les groupes dont la distance se trouve en dessous d'un seuil prédéterminée α. À chaque itération on détermine si les groupes D_i et D_j sont suffisamment proches, c'est-à-dire, si $Dist(D_i,D_j) \leq \alpha$. Dans le cas contraire, l'algorithme s'arrête et on garde les regroupements effectués.

Nous n'avons pas encore spécifié la façon de calculer la distance $dist(d_i,d_j)$. Celle-ci doit prendre deux vecteurs binaires et associer une valeur dans l'intervalle [0,1] qui quantifie la similitude entre d_i et d_j. À cet effet, nous proposons d'utiliser une nouvelle mesure dérivée du concept d'énergie textuelle (Fernández et al., 2007a).

3.3 L'énergie textuelle

Soit la matrice document-EL (2), où les valeurs x_{ji} représentent la présence ou l'absence du terme i dans le document j. Dans une configuration de réseau d'Hopfield (1982), les valeurs x_{ji} dans la matrice équivalent aux unités du réseau (Fernández et al., 2007, 2007a).

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & \ldots & \ldots & x_{np} \end{bmatrix} \quad (2)$$

L'énergie textuelle d'interaction entre les documents (Fernández et al., 2007) est calculée selon l'équation (3) :

$$E_{textuelle} = -\frac{1}{2}(X \times X^T)^2 \quad (3)$$

où les entrées (e_{ij}) de la matrice $E_{textuelle}$ sont toutes négatives ou nulles. Etant donné que l'objectif est de comparer la magnitude de la distance entre vecteurs binaires, sans perte de généralité, on peut considérer les valeurs absolues des entrées de la matrice :

$$E = \left| -E_{textuelle} \right| \quad (4)$$

Par sa symétrie, E peut être représenté comme un vecteur de distance énergétique (Molina, 2009). Le vecteur (5) contient la distance énergétique entre chaque paire de vecteurs binaires.

$$D_{ener} = [e_{12}, e_{13}, e_{14}, ..., e_{1n}, e_{23}, e_{24}, ..., e_{2n}, ..., e_{n-1n}] \quad (5)$$

Pour restreindre les valeurs des entrées dans l'intervalle [0,1], nous normalisons les entrées D_{ener} en divisant par le maximum.

Nous avons eu l'idée de combiner cette mesure avec le regroupement hiérarchique afin de créer des regroupements des définitions.

4 Le corpus de termes polysémiques en espagnol

Le CD est une structure discursive récemment étudiée et les corpus disponibles pour cette tâche sont rares voire inexistants. Nous avons eu donc besoin de créer notre propre corpus. Ses caractéristiques nécessaires pour les expériences sont très précises : d'un côté, il faut prendre en compte la quantité d'acceptions qu'un terme peut avoir selon son contexte (polysémie). D'un autre côté, il faut extraire d'Internet un nombre suffisant des données. Pour ces raisons, on a sélectionné minutieusement les termes à inclure dans le corpus des termes polysémiques en espagnol (CTPE). Selon les indications d'Alameda et Cuetos (1995), on a choisi d'abord une liste *L* de dix termes ayant les caractéristiques nécessaires : *L*={*aguja*, *barra*, *cabeza*, *casco*, *célula*, *golpe*, *punto*, *serie*, *tabla*, *ventana* (*aiguille, barre, tête, casque, cellule, coup, point, série, table*, *fenêtre*)}. Une fois la listé établit, le problème a été de trouver les CDs sur Internet. Ainsi on a associé la liste *L* à une liste de patrons définitoires afin de former un patron de recherche (PR). Un exemple de PR qui associe le terme *aguja* et le PD *ser+determinante* (*être + déterminant*) est : *la aguja es un (l'aiguille est un).* La figure 1 illustre quelques PRs avec le symbole <T> qui représente un terme générique.

la <T> est le	définit une <T>
la <T> est la	...
la <T> est un	nous définissons une <T>
les <T>s sont des	a défini la <T>
...	a défini une <T>
nous considérons la <T>	...

FIG. 1 – *Patrons de recherche.*

Nous avons utilisé les PRs avec un Web service (L'API BOSS de Yahoo!) afin d'extraire l'information nécessaire. La précision a été restreinte à cause des limitations du Web service. Il est possible -et très courant- de trouver des fragments textuels avec un PD mais qui ne sont pas de définitions. Par exemple, dans le fragment : *En general, el miedo a la aguja es el más frecuente,* (*En générale, la peur de l'aiguille est la plus fréquente*) on observe que le terme *aguja* est présent ainsi que le patron *ser+determinante*, mais évidemment il ne s'agit pas d'une définition. Un fragment de texte est un candidat à contexte définitoire (CCD) s'il contient un terme et un PD mais il n'est pas une définition (Sierra et Alarcón 2002). La très

grande quantité d'informations récupérées (environ 3,000 résultats par terme), s'opposait aux critères d'évaluation de l'étude (lecture et interprétation directe). La table 1 montre la quantité de CCDs extraits sur Internet pour chaque terme et genre de définition. À cause de la grande quantité de bruit obtenu nous avons décidé de réduire le nombre de termes. Les termes finalement retenus sont : *barra*, *célula*, *punto*, *ventana*.

	Analytiques	**Extensionnelles**	**Fonctionnelles**	
Barra (*Barre*)	1.863	307	467	**2.637**
Célula (*Cellule*)	5.352	649	533	**6.534**
Punto (*Point*)	1.702	422	750	**2.874**
Ventana (*Fenêtre*)	1.534	587	565	**2.686**
	10.451	**1.965**	**2.315**	**14.731**

TAB. 1 – *CCDs extraits du Web*.

5 Evaluation

5.1 Méthodologie d'évaluation

L'algorithme de regroupement a été exécuté pour chaque paire (terme-type de définition). Nous avons réalisé un balayage du seuil par distance de 0,01 jusqu'à 1,00, avec un pas de 0,01. Nous avons affiché uniquement les groupes réunissant au moins deux définitions, en éliminant ceux possédant une seule. Afin que le lecteur puisse observer toutes les définitions originellement présentes lors des expériences, nous avons inclus le groupe absolu dont la valeur du seuil par distance est de 1,00.

L'analyse qualitative a consisté à une lecture et interprétation directe des groupes générés, effectuée par un évaluateur humain. Les résultats peuvent être consultés sur le site web http://saussure.iingen.unam.mx/~amolinav/resultados

Afin de comparer les résultats avec ceux de la distance énergétique, nous avons utilisé une adaptation de la distance de Hamming (1950) comme *baseline*. Nous avons divisé le nombre d'entrées différentes entre la longueur des vecteurs afin d'obtenir des valeurs dans la plage [0,1]. Nous présentons les courbes en utilisant les deux critères de distance : l'énergie textuelle et la distance de Hamming.

Nous avons évalué la qualité des regroupements avec trois mesures : le nombre de groupes, la précision et le rappel. À notre connaissance, il n'existe pas une mesure unifiant les trois critères. Même si la *F*-mesure combine la précision et le rappel (Van Rijsbergen, 1979), nous avons décidé de les isoler, car l'adaptation de la précision définie ci-dessous implique la lecture et une interprétation humaine. Le calcul du rappel étant complètement automatique, leur combinaison pourrait donc fausser les résultats.

5.2 Nombre de groupes

Après avoir utilisé la distance énergétique, le nombre de groupes augmente en proportion de la valeur de seuil par distance. Le type de courbe est pratiquement le même pour tous les

types de définition. Nous avons déduit que le nombre de groupes générés est indépendant du type de définition. La figure 2 illustre le comportement du nombre de groupes en fonction de la valeur de seuil par distance pour les définitions extensionnelles.

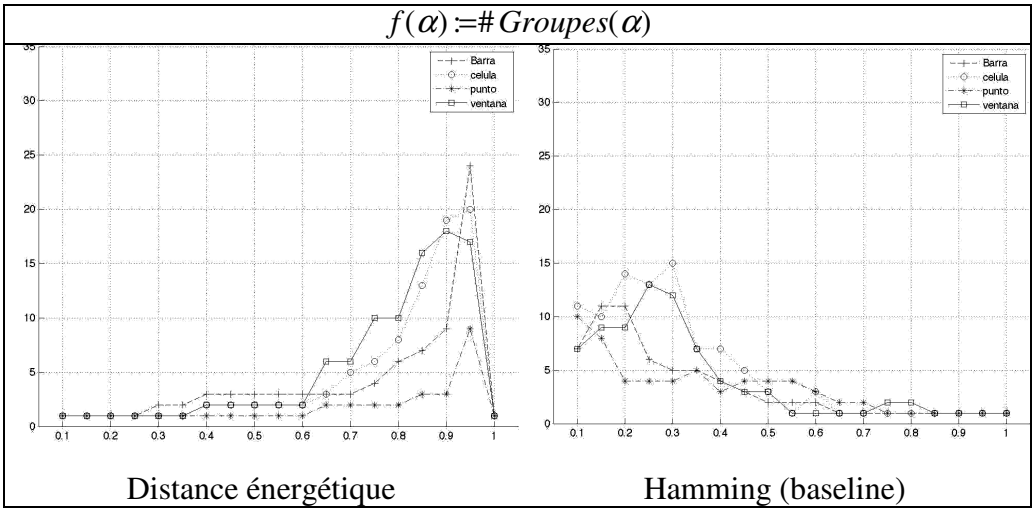


FIG. 2 – *Nombre de groupes en fonction de la valeur de seuil par distance pour les définitions extensionnelles.*

5.3 Evaluation du rappel et de la précision

Dans la recherche d'information, le rappel est la proportion de documents récupérés dans une recherche. Nous avons adapté cette définition de rappel comme la proportion de définitions intégrées à un groupe par rapport au total des définitions. C'est-à-dire, combien de définitions nous réussissons à intégrer dans un regroupement. Le rappel prendra la valeur 0 si aucun groupe, avec au moins deux définitions, n'a été identifié ; et 1 si toutes les définitions ont été intégrées dans un groupe. La formule utilisée pour calculer le rappel est la suivante :

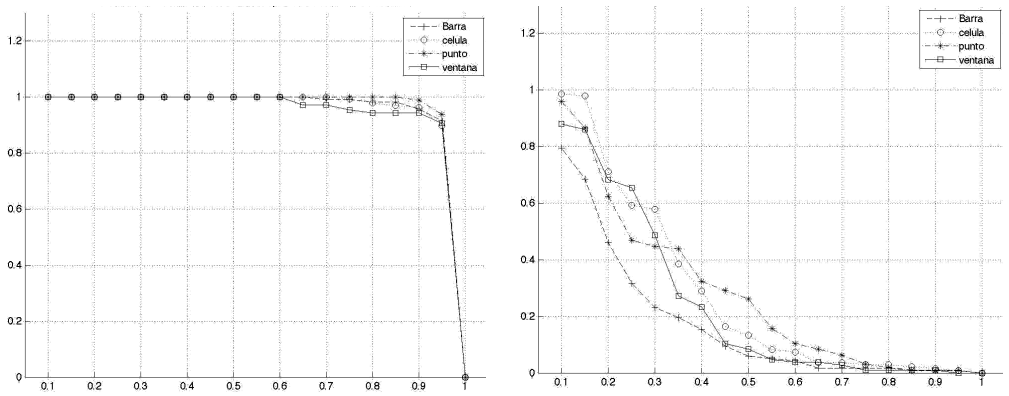
$$r = \frac{\left|\{Totalité\_de\_définitions\} \cap \{Définitions\_inclues\_dans\_un\_groupe\}\right|}{\left|\{Totalité\_de\_définitions\}\right|} \qquad (6)$$

La précision est définie comme la fraction de documents pertinents parmi les documents récupérés dans une recherche. Nous avons aussi adaptée cette définition pour quantifier la proportion d'intrus dans le regroupement généré. Elle indique combien d'erreurs nous avons commises après avoir généré les groupes. La précision obtient valeur de 0 si aucun groupe définit auparavant est formé, et de 1 si aucun groupe contient d'intrus. La formule utilisée pour la précision est :

$$p = \frac{\left|\{Définitions\_inclues\_dans\_un\_groupe\} - \{Intrus\}\right|}{\left|\{Définitions\_inclues\_dans\_un\_groupe\}\right|} \qquad (7)$$

Le comportement de la précision et du rappel est indépendant du type de définition. Pour des raisons de clarté nous montrons seulement les résultats des définitions extensionnelles. La figure 3 illustre le comportement de la précision en fonction de la valeur du seuil par distance et la 4 illustre celui du rappel.

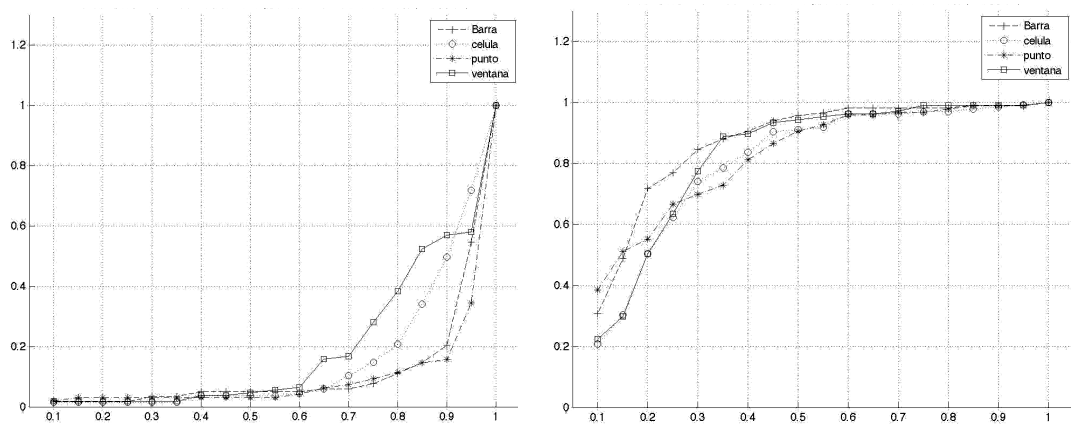
$f(\alpha) \coloneqq précision(\alpha)$



Distance énergétique Hamming (*baseline*)

FIG. 3 – *Précision en fonction de la valeur du seuil par distance pour les définitions extensionnelles.*

$f(\alpha) \coloneqq Rappel(\alpha)$



Distance énergétique Hamming (*baseline*)

FIG. 4 – *Rappel en fonction de la valeur du seuil par distance pour les définitions extensionnelles.*

5.4 Une analyse qualitative

En général, on valide des regroupements qui reflètent les différentes acceptions d'un terme, mais aussi, on distingue des subtilités existantes dans les définitions même à l'intérieur d'une acception identique. Nous pouvons affirmer que l'algorithme regroupe des définitions minutieusement, car il réussit à regrouper des définitions structurellement distinctes mais avec une signification équivalente. L'analyse qualitative semble ainsi très encourageante. Considérez comme exemple de ce phénomène les deux définitions suivantes de *célula* (cellule):

La célula se compone de un núcleo envuelto en protoplasma, alrededor del cual hay una membrana que separa la célula de su medio ambiente[4].

La célula consta de una membrana celular que envuelve una masa viscosa y granulosa llamada protoplasma, en la cual se encuentran todos los organelos celulares, incluido el núcleo[5].

Dans la première, nous pouvons déduire que la cellule est composée d'un noyau qui est enveloppé d'un protoplasme à son tour enveloppé d'une membrane. La deuxième définition est structurellement inversée, mais sémantiquement équivalente : la cellule est composée d'une membrane qui enveloppe le protoplasme où l'on trouve le noyau.

Un autre résultat important de l'analyse est le suivant : à mesure que la valeur de seuil tend vers 1, les groupes deviennent plus spécifiques et parfois inconvenablement explicites dans la signification des mots qui les conforment. Nous avons aussi observé que lorsque plus de définitions sont incluses dans le regroupement, un bruit fort s'introduit au sein de groupes qui ont été crées. Nous parlons des intrus qui apparaissent dans les groupes et qui sont incongrus avec la majorité des définitions du groupe.

5.5 Analyse quantitative

Nous avons constaté que le comportement de l'algorithme est indépendant du type de définition. Cela montre que l'algorithme peut être utilisé pour n'importe quel type de texte et pas seulement pour des définitions.

Le nombre de groupes, la précision et le rappel sont proportionnels au seuil. Il faut mentionner que, mis à part le type de définition, le comportement de l'algorithme peut être divisé en zones par rapport à la valeur de seuil par distance :

1. La zone 1 où $0 \leq \alpha \leq 0.7$: dans cette zone on obtient une très bonne précision (supérieure au 90%) et un rappel bas (inférieur à 40%). Il faut utiliser une valeur α dans cet intervalle pour obtenir des acceptions courantes, un nombre réduit de groupes (environ 5) avec peu de définitions et sans la présence d'intrus dans les groupes générés.

[4] La cellule est composée d'un noyau enveloppé dans le protoplasme, autour duquel il y a une membrane qui sépare la cellule de son milieu.

[5] La cellule est composée d'une membrane cellulaire qui enveloppe une masse visqueuse et granuleuse appelée le protoplasme, dans lequel on trouve toutes les organelles cellulaires, y compris le noyau.

2. La zone 2 où $0.75 \leq \alpha \leq 0.85$ se caractérise par un haut degré de précision (autour de 80%) et un rappel moyen (autour de 50%). Cet intervalle maintient un bon équilibre entre la précision et le nombre de groupes générés (environ 10) ainsi qu'un rappel acceptable.
3. La zone 3 où $0.85 \leq \alpha \leq 0.99$, obtient une précision moyenne (environ 50%) et un rappel élevé (environ 80%). Le nombre de groupes générés dans cette zone est trop élevé (supérieur à 20) -il y a plus de groupes que d'acceptions-, par contre chaque groupe est très précis en termes de la cohérence de sa signification.

6 Conclusions et travail futur

Même si ce travail a été testé sur un corpus de définitions en espagnol, la méthode présentée est indépendante des considérations linguistiques. Une étude d'adaptation à la langue française est en cours au Laboratoire Informatique d'Avignon. En ce qui concerne la complexité, il faut considérer l'espace mémoire (en raison de la génération de la matrice document-EL qui est toujours d'ordre $O(P^2)$, où P est le nombre de documents. Ainsi, la méthode est bien adaptée pour regrouper des textes courts (quelques dizaines de phrases). Par exemple des extraits de résultats des moteurs de recherche (*snippets*), les manchettes, les résumés automatiques, etc.

Une amélioration possible de notre méthode de regroupement de définitions consiste à calculer dynamiquement la valeur de seuil grâce à laquelle on obtiendrait le meilleur regroupement en prenant en compte les caractéristiques désirées (nombre de groupes précis, haute précision, rappel élevé, etc.). On pourrait aussi fixer la valeur du seuil α afin de comparer des distances binaires différentes. Ces études son actuellement en cours.

Remerciements

Ce travail a été partiellement financé par le Conacyt (Mexique) pour la bourse de A. Molina et aussi par la collaboration entre le *Grupo de Ingénieria Linguistica* (GIL/UNAM Mexico) et le Laboratoire Informatique d'Avignon (LIA/UAPV, France).

Références

Aguilar, C. et Sierra, G. (2009). A formal scope on the relations between definitions and verbal predications. *1st International Workshop on Definition Extraction*. Borovets, Bulgaria, 1-6.

Alameda, J. et Cuetos, F. (1995). *Diccionario de Frecuencia de las unidades lingüísticas del castellano* : Servicio de Publicaciones, Universidad de Oviedo.

Alarcón, R., Sierra G. et Bach, C. (2009). Description and Evaluation of Definition Extraction System for Spanish language. *1st International Workshop on Definition Extraction*, Bulgaria, 7-13.

Fernández, S., SanJuan, E. et Torres-Moreno, J-M. (2007). Textual Energy of Associative Memories: performants applications of ENERTEX algorithm in Text summarization and topic segmentation. *Proceedings of MICAI 2007*, 861-871.

Fernández, S., SanJuan, E. et Torres-Moreno, J-M. (2007a). Energie textuelle de mémoires associatives. *Conference TALN 2007*, 25-34.

Hamming R. (1950). Error detecting et error correcting codes. *Bell System Technical Journal*, 2: 147–160.

Hopfield J. (1982). Neural Networks and Physical Systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA*, 9: 2554–2558.

Malaisé, V. (2005). *Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles á partir de corpus textuels*. Thèse de doctorat, Université Paris 7–Denis Diderot.

Molina, A. (2009). *Agrupamiento semántico de contextos definitorios*. Rapport de master, Universidad Nacional Autónoma de México.

Rebeyrolle, J. et Ludovic, T. (2000). Repérage automatique de structures linguistiques en corpus: le cas des énoncés définitoire. *Cahiers de Grammaire,* 25:153-174.

Rodríguez, C. (2004). *Metalinguistic Information Extraction from Specialized Texts to Enrich Computational Lexicons*. Thèse de doctorat, Universidad Pompeu Fabra, Barcelona.

Salton, G. (1971). *The SMART Retrieval System: Experiments in automatic document procesing*. Prentice Hall, Englewood Cliffs.

Sierra, G., Alarcón, R. (2002). Identification of recurrent patterns to extract definitory contexts. *Lecture Notes in Computer Science*. Springer-Verlag, 2276 : 436-438.

Sierra, G., Alarcón, R., Medina, A., Aguilar, C.A. (2004). Definitional contexts extraction from specialised texts. *Practical Applications in Language and Computers*, édité par Barbara Lewandowska.Tomaszczyk. Frankfurt: Peter Lang. 21-31.

Sierra, G., Alarcón, R., Molina, A. et Aldana, E. (2009). Web Exploitation for Definition Extraction, en *Proceedings of LA-WEB 2009 (7th Latin American Web Congress).*

Sorensen, T. (1948). A method of estimating groups of equal amplitude. *Plant sociology based on similarity of species content*, 5: 1-34.

Van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth (2ème édition).