UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS
Département d'informatique et d'ingénierie

Impact de l'intégration d'une ontologie normée XBRL à la classification automatique de textes : Une application aux nouvelles financières

Par

Sadia Messaoudi

Département d'informatique et d'ingénierie

Mémoire pour l'obtention du grade de
maître ès sciences (M. Sc.)

13 MAI 2011

Remerciements

Je tiens à remercier tous ceux qui m’ont aidée à franchir ce cap malgré toutes les difficultés de la vie courante.

Je tiens aussi à rendre hommage à ceux qui ont cru en mes capacités malgré les courants contraires qui ont tout fait pour m’arrêter. Pour cela, je tiens à remercier particulièrement Alain pour sa présence constante et son aide indispensable ainsi que pour son amitié inestimable.

Je remercie aussi Stéphane de m’avoir offert l’occasion de mesurer mes capacités et de connaître mes forces et mes faiblesses sachant que la vie est faite de défis et de combats.

J’aimerais aussi envoyer un clin d’œil à mon âme sœur qui a su résister aux aléas de mes travaux infinis.

Une place sans frontières est spécialement réservée pour ma petite fille, sans laquelle, ce rêve ne se serait jamais réalisé. Sa présence dans ma vie m’a donné la force de me battre et d’atteindre mes buts qui jusqu’alors me paraissaient insurmontables. Je remercie sincèrement celui qui m’a fait don de ce cadeau inestimable.

À tous ceux qui ont été présents de près et de loin, je tiens à dédier ce mémoire.

Résumé

Bien que beaucoup de méthodes développées dans le domaine de la catégorisation automatique de textes (CAT) ont permis d'atteindre des niveaux de précision appréciables lorsqu'il s'agit de textes à structure simple (ex. courriels, résumés, etc.), il reste néanmoins encore des défis à relever dans le cas de documents complexes tels les nouvelles financières et autres analyses similaires à base de connaissance. Cette complexité rend plus difficile la formalisation et la mise à jour d'une base de connaissance représentative, ce qui influencera directement la fouille de textes dans le repérage de sujets communs entre les textes et les composantes (par analyse de similarités et de hiérarchies) et leur suivi à travers le temps (ex. *Topic Detection and Tracking*).

Dans ce mémoire, nous proposons d'adopter, comme modèle de représentation formelle des connaissances, les ontologies normées qui ont récemment démontré une amélioration dans les résultats de classification. Parmi les recherches réalisées dans ce domaine, nous pouvons citer l'ontologie Wikipedia qui possède à elle seule, en 2007, 2 millions d'entrées [1], la classification multilingue à base d'ontologies [2] et l'intégration des ontologies dans les tâches de recherche d'information (spécialement dans le regroupement de textes et les tâches de classification) [3]. Afin de valider notre approche, des expériences seront menées via l'utilisation du classificateur commercial IBM Classification Module (ICM, un module d'IBM OmniFind). Nos tests de classification seront effectués sur un sous-ensemble précis de nouvelles du secteur financier provenant du *Reuters Corpus Version 1* (RCV1) lequel, avec ses 810,000 nouvelles, correspond à la plus large collection de dépêches disponibles.

Abstract

Despite the fact that many methods, developed in the field of automatic text categorization, have achieved significant levels of precision when it comes to simple structure of texts (e.g. emails, summaries, etc.). Nevertheless, there remains much to do in the case of complex documents such as financial news and similar knowledge-based analyses. This complexity makes it more difficult to formalize and update a representative knowledge base, which directly influence text mining in the identification of common issues between the text and the components (by analysis of similarities and hierarchies) and there monitoring through time (e.g., Topic Detection and Tracking).

In this research, we propose to adopt, as a model for formal representation of knowledge, normalized ontology which has recently demonstrated an improvement in classification results. Among the research conducted in this area we include Wikipedia ontology that contains, in 2007, two million entries by itself [1], the multilingual classification based on ontology [2] and the integration of the ontology inside information retrieval tasks (especially in the grouping of texts and tasks of classification) [3]. To validate our approach, experiments will be conducted through the use of a commercial classifier IBM Classification Module (ICM, a module of IBM OmniFind). Our classification tests are performed on a specific subset of new financial sector from the *Reuters Corpus Version 1* (RCV1) which, with its 810 000 news, is considered as the largest collection of news available.

TABLE DES MATIÈRES

Liste des figures

Liste des tables

Liste des abréviations

ACE	*Augmentation de la compatibilité vers l'enrichissement*
AP	*Amélioration partielle*
AT	*Amélioration totale*
CAT	C*atégorisation automatique de textes*
CSV	*Comma-separated values*
DARPA	*Defense Advanced Research Projects Agency*
DP	*Diminution partielle*
DT	*Diminution totale*
EA	*Enrichissement augmenté*
ECM	*Enterprise Content Management*
ED	*Enrichissement diminué*
ES	*Enrichissement stable*
FN	*False Negative*
FP	*False Positive*
hF	*Mesure hiérarchique de la F-mesure*
hP	*Mesure hiérarchique de la Précision*
hR	*Mesure hiérarchique du rappel*
IAS	*International Accounting Standard*
ICM	*IBM Classification Module*
IFRS	*International Financial Reporting Standards*
LH	*Liste hiérarchique*
LN	*Liste normée*
LS	*Liste simple*
NBC	*Nouvelles bien classées*
NBCPE	*Nouvelles bien classées avec perte d'enrichissement*
NIST	*National Institute of Standards and Technology*
NLP	*Natural Language Processing*
NMC	*Nouvelles mal classées*
NPBC	*Nouvelles partiellement bien classées*
PCGR	*Principes Comptables Généralement Reconnus*
RCV1	*Reuters Corpus Version 1*
RME	*Relationship Modeling Engine*
SVM	S*upport vector machines*
TN	*True Negative*
TP	*True Positive*
TREC	*Text REtrieval Conference*
XBRL	*eXtensible Business Reporting Language*
XML	*eXtensible Markup Language*

CHAPITRE 1

Introduction

1.1 Champs d'étude

Le présent projet s'inscrit dans le cadre de la recherche d'information (*Information Retrieval*) en général, et dans le domaine de la CAT en particulier.

Dans la classification, un expert spécifie des catégories principales afin de générer des règles qui serviront à classifier de nouveaux documents [1]. Dans la classification de textes en général, on associe des catégories à des documents à travers une tâche de tri automatique. Cette tâche est actuellement à la croisée des chemins entre la recherche d'information et l'apprentissage machine [2].

1.1.1 Traitement automatique du langage naturel

Le traitement naturel des langues (*Natural Language Processing* (NLP)) concerne les domaines ayant pour objectif principal l'analyse et le traitement de tous les aspects du langage humain. Cette discipline touche particulièrement notre étude car, étant intégrée au noyau des classificateurs de documents, elle leur permet d'analyser la sémantique des textes afin d'assigner, à ces derniers, des catégories correspondantes. Le fonctionnement de la NLP sera présenté de façon plus détaillée dans l'annexe A présentant le fonctionnement interne du moteur de classification de l'engin de classification ICM.

1.1.2 Classification automatique de textes

La CAT est souvent requise en gestion des documents numériques, en particulier la classification hiérarchique selon une taxonomie (ou ontologie) du domaine d'application. Dans ce mémoire, nous proposons d'améliorer la performance de ce type de classification via l'utilisation d'une ontologie normée. L'annexe B présente plus en détail la classification en général et la CAT en particulier**.**

1.1.3 Ontologie et normalisation

Dans le but d'améliorer les performances de la classification automatique de documents, nous avons adopté un modèle de représentation formelle de connaissances, à savoir les ontologies normées qui ont récemment démontré une nette amélioration dans les résultats de la classification (voir l'annexe D pour avoir plus de détails sur les ontologies).

Comme un domaine de connaissances ne peut être conceptualisé que dans un contexte d'usage précis [3], l'ontologie retenue pour cette étude sera basée sur les normes d'un domaine d'application, à savoir l'analyse des marchés financiers (voir section K.3).

1.2 Structure du mémoire

Dans ce qui suit, nous décrirons, au premier chapitre le champ d'étude dans lequel se situe le projet actuel en présentant l'impact du domaine de la classification automatique de textes dans nos travaux ainsi que les objectifs visés dans les recherches effectuées. Ces derniers seront détaillés dans le deuxième chapitre dans lequel nous discuterons des objectifs théoriques et pratiques issus des travaux de recherche dont le cadre théorique et la méthodologie suivis seront détaillés respectivement au chapitre trois et quatre. Le cadre théorique contiendra, entre autres, la liste des mesures de performance utilisées pour le calcul des performances de notre classificateur. La méthodologie appuiera, quand à elle, les étapes nécessaires à la préparation de l'environnement de classification pour les tests et présentera un sommaire final des résultats obtenus. Le cinquième chapitre servira à conclure à partir de tous les constats notés dans le chapitre des résultats tout en présentant les apports scientifiques ainsi que les ouvertures de recherche futures. Enfin, les détails techniques seront élaborés au niveau des annexes qui serviront comme information supplémentaire appuyant le projet présenté et agrémentant l'information scientifique en approfondissant les notions succinctement présentées dans les chapitres précédents.

CHAPITRE 2

Objectifs

2.1 Objectifs principaux de recherche

Le principal objectif de cette étude est l'amélioration des performances de la CAT en général. De façon plus particulière, on aimerait prouver que cette amélioration peut être atteinte grâce à l'utilisation d'une taxonomie développée à partir d'une ontologie normée versus deux (2) autres types de taxonomies, soient plate et hiérarchique non-normée.

Étant donné que cette étude concerne particulièrement l'amélioration du traitement de l'information financière, nous nous concentrons sur l'évaluation de l'utilisation d'une ontologie normée dans le cadre de la classification hiérarchique des nouvelles financières.

Ainsi on a procédé à la codification des nouvelles financières selon une structure normée afin de mieux classifier les nouvelles selon les variables comptables sous-jacentes. Cette classification simulée a été réalisée grâce à l'utilisation d'un classificateur commercial IBM Classification Module v.8.6 (Voir Annexe A) et de la base de données Reuter Corpus Volume 1 (RCV1) publiée en 2002, ses 800 000+ nouvelles, étalées de septembre 1996 à août 1997, ont été classifiées manuellement dans une hiérarchie à 2 niveaux. Nous utilisons seulement les nouvelles liées aux *fusions* et *acquisitions*, soit environ 40 000+ nouvelles.

2.2 Justifications pour une ontologie normée

Le choix d'une liste normée pour améliorer la CAT s'appuie sur les principales justifications suivantes :

1. La norme couvrira le domaine de manière plus complète.
2. L'hiérarchie des termes sera plus utile à leur interprétation.
3. La reproduction de l'étude sera plus aisée.
4. L'application sera plus facile à intégrer dans un système.
5. Les diverses propriétés (ontologie) de la liste normée impliquent un certain nombre de règles de décision ou logiques d'affaire que les autres listes non-normées n'ont pas.

2.3 Avancées visées

2.3.1 Introduction

Des techniques de forage de données ont été appliquées dans beaucoup de domaines tels la médecine et l'informatique. Des améliorations ont été observées dans les résultats obtenus grâce aux méthodes constamment révisées. Au même moment, le domaine des finances accuse un certain retard dans l'automatisation de la collecte et du traitement de l'information financière dont le volume ne cesse d'augmenter chaque année. Cela va à l'encontre des besoins de traiter l'information dans des délais très courts. Les besoins et aussi les scandales financiers actuels ont provoqué la création de nouvelles lois et réglementations comptables en vue d'obtenir un *reporting* financier beaucoup plus rigoureux et beaucoup plus volumineux [4].

Tous ces besoins ne nécessitent pas seulement un standard comptable pour le formatage de l'information financière afin de simplifier le travail d'analyse, mais aussi et surtout la normalisation de l'échange des informations. En effet, des études récentes semblent indiquer que les normes comptables auraient un impact sur la lecture et le traitement des nouvelles financières (surtout celles liées aux *fusions* et *acquisitions* adoptées dans le présent projet).

Pour cette raison, nous supposons que l'utilisation d'une ontologie normée aidera à améliorer la classification hiérarchique de nouvelles financières.

2.3.2 Classification de nouvelles financières complexes

En codant les nouvelles financières selon qu'elles discutent d'un ou plusieurs sujets selon une structure normée, nous visons à mieux classifier les nouvelles selon les variables comptables sous-jacentes. Nous pourrons ainsi améliorer l'analyse stratégique des transactions de *fusion* et *acquisitions* rapportées dans les nouvelles, surtout en identifiant la source précise dans les états financiers de l'information traitée, ainsi que les concepts auxquels ces données comptables sont associées.

2.3.3 Amélioration de la classification par l'intégration d'une ontologie normée

L'utilisation d'une ontologie normée dans le domaine des finances en particulier, permettra de profiter des normes comptables qui semblent avoir un impact direct sur le traitement des nouvelles financières. En effet, ces dernières (normes comptables) simplifient l'analyse et la normalisation de l'échange d'information.

2.3.4 Utilisation de la norme XBRL pour représenter l'information financière

Afin d'obtenir une information financière sous un format facilement analysable et instantanément catégorisable par les systèmes et organisations récepteurs sans perdre l'exactitude des données d'origine, nous adopterons le format XBRL (*eXtensible Business Reporting Language*) qui est un langage de *reporting* libre de droits fondé sur le standard XML (*eXtensible Markup Language*) (voir l'annexe E).

En effet, le XBRL représente une solution fiable pour assurer la collecte et la publication d'information financière, le partage et l'analyse de cette dernière, l'indépendance des données par rapport aux applications d'origine et enfin le multilinguisme dans la traduction des libellés des taxonomies [5]. XBRL est présentement à l'essai au Canada pour formaliser la soumission des rapports annuels à la base SEDAR[1]. XBRL touche l'information commerciale en particulier et permet l'extraction automatique de l'information.

[1] « *System for Electronic Document Analysis and Retrieval (SEDAR*) est une base de données à laquelle, en 2008, toutes les sociétés par actions et tous les fonds d'investissement canadiens doivent être inscrits. SEDAR est opéré par le *Canadian Securities Administrators* qui regroupe les agences de niveaux provincial et territorial et qui contrôlent le fonctionnement des bourses » Wikipédia 8 jan 2009 (voir lien : http://fr.wikipedia.org/wiki/SEDAR).

L'éditeur d'ontologies choisi (Protégé de Stanford) va permettre, dans cette étude, d'importer et d'augmenter diverses éditions de la norme XBRL en respect, entre autres, avec celle adoptée par le comité XBRL Canada, le Cadre canadien d'information financière (CCIF).

2.3.4.1 Schéma XBRL choisi pour la construction de la liste

Pour construire l'ontologie normée du domaine, nous utiliserons le schéma du XBRL v.2.1, selon le *International Financial Reporting Standards* (IFRS)[2] [6]. Nous nous attarderons en particulier sur 2 normes clés (voir annexe F) :

1. *International Accounting Standard 1* (IAS 1) pour la présentation des états financiers :
   - 1.1. [310005] État des revenus - En fonction des dépenses - États financiers distincts (*Income statement, by function of expense - Separate financial statements*)
   - 1.2. [220005] État de la situation financière - Ordre de liquidité - États financiers distincts (*Statement of financial position, order of liquidity - Separate financial statements*)
2. IFRS 3 pour les Notes aux états financiers pour les regroupements d'entreprises :
   - 2.1. [817000] Notes – Regroupement d'affaires (*Business combinations*)

La hiérarchie offerte par ces 2 normes sera notre point de concentration, et non les mots clés ou expressions en soit. C'est la structure logique du rapport annuel qui est surtout normée ici, et non pas les concepts, car ceux-ci sont issus des normes comptables internationales, et sont donc assez communs dans les nouvelles.

2.3.4.2 Traductions comptable de la norme IAS 22 et de l'IFRS 3

La norme IFRS 3, dont la plus récente révision date de 2008 et a pris effet le 1er juillet 2009, est en fait une version modifiée de la norme IAS 22[3][7]. Elle l'a remplacé le 1er janvier 2005. Celle-ci avait été éditée en 1993 et avait pris effet le 1er janvier 1995 [8]. Puisque les nouvelles du RCV1 sont datées de 1996-1997, les rapports annuels et commentaires des analystes font usage de la norme IAS 22. Pour assurer la cohérence entre les rapports de l'époque et notre utilisation de la norme IFRS 3 sous XBRL édition 2008, nous avons fait la comparaison des concepts et de la hiérarchie des deux normes, selon les suggestions d'une analyse menée par KPMG [9].

---

[2] « Les normes internationales d'information financière, plus connues au sein de la profession comptable et financière sous leur nom anglais de International Financial Reporting Standards ou IFRS sont des normes comptables, élaborées par le Bureau des standards comptables internationaux destinées aux entreprises cotées ou faisant appel à des investisseurs afin d'harmoniser la présentation et la clarté de leurs états financiers. » Wikipédia du 20 juillet 2009. Lien : http://fr.wikipedia.org/wiki/IFRS

[3] « *Le changement majeur de l'IFRS 3 par rapport à l'IAS 22 concerne les modalités de valorisation du goodwill issu d'un regroupement d'entreprises. L'IAS 22 prévoyait un amortissement systématique du goodwill sur sa durée de vie économique, celle-ci ne pouvant pas excéder 20 ans. Pour IFRS 3, le goodwill doit être maintenu au coût d'acquisition, déduction faite des pertes de valeurs. Ainsi le goodwill n'est plus amorti, mais il doit faire l'objet d'un test de perte de valeur au moins une fois par an ou plus fréquemment, si des événements ou changements de circonstances indiquent qu'une perte de valeur pourrait exister.». (« Impacts de la mise en place des normes IFRS sur les capitaux propres », ,Bazire, S. et M.-N Maffon de 2005, p 119).*

Nous réorganisons les concepts issus de la liste hiérarchique selon la structure normée des concepts comptables et des combinaisons d'entreprises. Nous retrouverons, en annexe G, les sujets liés à la structure et aux raisons dans les *Notes aux états financiers*, et les sujets liés aux résultats dans les deux états financiers que sont l'*État des résultats* (*Income Statement*) et *Bilan* (*Statement of Financial Position*).

CHAPITRE 3

Cadre théorique

3.1 Introduction

Dans ce troisième chapitre, nous allons présenter la méthode utilisée dans ce projet pour le calcul des performances de notre système automatique de classification. On détaillera notamment les mesures de Sébastiani pour les classificateurs binaires, les mesures hiérarchiques pour les classificateurs hiérarchiques, les mesures de distance pour les classificateurs hiérarchiques utilisant des classes parentes et enfin une mesure proposée afin de palier aux lacunes des autres mesures dans le calcul détaillé de la qualité de la classification.

3.2 Critères d'évaluation des résultats d'une classification non plate

La performance d'un système de recherche d'information est calculée selon la pertinence de ses résultats à travers des mesures d'efficacité telles que la *précision* et le *rappel*. Ces deux mesures qui sont reliées, parfois de façon conflictuelle (voir la Figure 23 de l'Annexe C), sont plus difficiles à exploiter lorsque le corpus utilisé est assez grand. Aussi, elles ignorent l'interactivité avec l'utilisateur. Pour cette raison, nous avons utilisé des mesures de performance destinées à la classification non hiérarchique, tout en les complétant par des mesures qui prennent en compte la performance du parent/enfant dans la hiérarchie [17]. Nous avons surtout appliqué les critères utilisés par Ceci & Malerba ([11], p.57) pour l'évaluation de la classification par hiérarchie :

1. *Y a-t-il amélioration de performance lorsque, comparé à un classificateur plat, le classificateur hiérarchique est construit dans le cadre proposé?*
2. *Est-ce que le cadre proposé minimise la distance (en terme d'arborescence) entre la classe correcte et celle retournée lorsque le document n'est pas bien classifié?*
3. *Est-ce que le cadre proposé améliore actuellement l'efficacité de calcul des algorithmes d'apprentissage?*
4. *Quelle stratégie de sélection de caractéristique est la plus prometteuse pour la catégorisation hiérarchique?*
5. *Quel classificateur a la meilleure performance dans le cadre proposé?*

Pour plus de détails sur la logique d'utilisation des mesures hiérarchiques, consulter l'annexe B, partie B.6.

3.3 Défis pour l'amélioration d'une classification hiérarchique

Bien qu'elle se soit initialement inspirée des méthodes d'évaluation des classificateurs non hiérarchiques (voir la méthode détaillée sous l'annexe C), la méthode d'évaluation des classificateurs hiérarchiques obéit à des contraintes différentes. Ainsi, avant d'évaluer les résultats de la classification automatique des nouvelles, il faut s'attendre à affronter plusieurs défis afin d'améliorer les performances de cette dernière [11] :

1. Il faut généralement assurer que la classification d'un texte à un niveau détaillé d'une classe soit consistante avec sa classification dans une classe supérieure [12].
2. Nous devons être capables d'améliorer la taxonomie par apprentissage, et possiblement générer automatiquement une hiérarchie consistante à tous les niveaux [13].
3. Nous nous préoccupons également de la performance du classificateur selon les divers degrés de formalisme dans les textes, en particulier s'ils sont écrits selon un vocabulaire propre à la hiérarchie [17].
4. Nous pouvons induire le formalisme des documents en appliquant des étiquettes (*Part of Speech* ou POS tags) reliées à une ontologie proche de la hiérarchie ciblée [14].
5. Si ces efforts ne réduisent pas le nombre de classes choisies dans la hiérarchie, et n'améliorent pas la consistance de la classification, nous pouvons procéder par itérations pour ordonner les classes ayant un score similaire, et ainsi choisir la classe la plus probable [15].
6. Enfin, nous pouvons améliorer la performance de la classification en faisant usage conjointement d'une hiérarchie normalisée ainsi que d'un ensemble de classes suggérées par un groupe d'utilisateurs [16].

3.4 Choix des mesures

Étant donné que cette étude concerne de près la classification automatique hiérarchique multi-classes, il est très important d'utiliser des mesures de calcul de performance orientées vers la classification hiérarchique en particulier.

Les mesures appliquées aux classificateurs plats telles la précision et le rappel, ne conviennent pas à une classification hiérarchique dans la forme qu'on leur donne car elles ne prennent pas en considération les types d'erreurs liées à la mauvaise classification [12].

Ainsi, si une classe choisie est proche (parent ou frère) de la classe vraie, alors l'erreur est moins importante que si cette dernière est distante. Pour cette raison, la meilleure façon d'évaluer ce genre de classification est d'adopter une mesure hiérarchique qui respecte les trois conditions décrites dans l'étude réalisée par Kiritchenko & al. 2006 [12].

Considérons l'exemple suivant d'un arbre de classes appartenant à une classification hiérarchique. (G est supposée être la classe vraie d'un document donné.) :

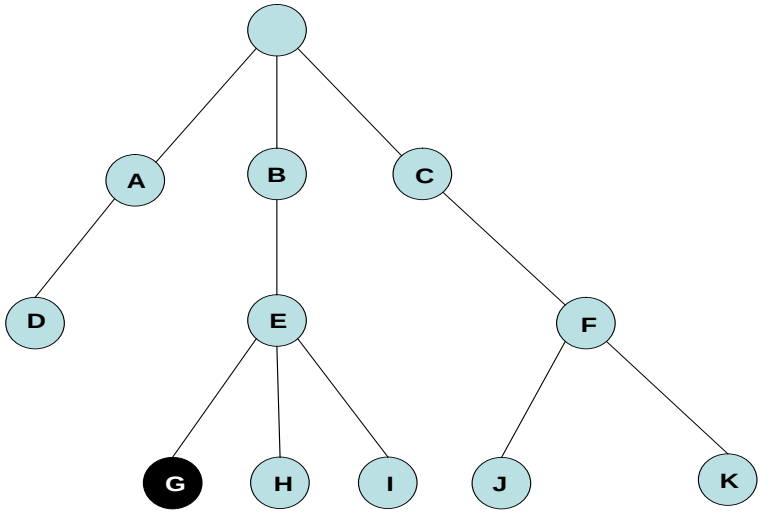


Figure 1: Une hiérarchie de classes de type DAG

Les conditions proposées par Kiritchenko & al. sont les suivantes :

1. La mesure doit prendre en considération la classification partiellement correcte.

   Exemple
   En se basant sur la Figure 1, la classification d'un document sous la classe I devrait être moins pénalisante que celle sous la classe D en considérant G comme la classe vraie.

2. La mesure doit noter différemment les erreurs de distance selon le type de distance reliant la classe vraie de celle obtenue.

   Exemple
   En se basant sur la Figure 1, la classification d'un document sous la classe E devrait être moins pénalisante que celle sous la classe B en considérant G comme la classe vraie.
   « La mesure doit prendre en considération le nœud le plus proche selon son niveau »

3. La mesure doit pénaliser la mauvaise classification à un niveau supérieur.

   Exemple
   En se basant sur la Figure 1, et en considérant 2 cas de figure suivants :
   1. G est la classe vraie et I est la classe choisie par le classificateur
   2. A est la classe vraie et C est la classe choisie par le classificateur.

   Alors, la classification d'un document selon le premier cas devrait être moins pénalisante que si la classification était selon le 2ème cas.

Ainsi, afin de respecter ces 3 conditions, les mesures doivent prendre en considération les ancêtres et les descendants d'une classe en plus de pouvoir mesurer la distance.

3.5 Mesures traditionnelles

3.5.1 Mesures issues du tableau de contingence de Sébastiani pour la classification plate

Les mesures initialement utilisées sont celles que Sébastiani, entre autres auteurs, a adopté pour présenter les moyens mathématiques de calcul des performances d'un classificateur automatique binaire. Ces mesures permettent d'avoir une idée générale des niveaux de performance de notre système de classification :

► La *précision* (degré de certitude) représente la probabilité conditionnelle que si un document d_j, pris au hasard, est classifié sous une catégorie c_i alors cette décision est correcte [29] :

Précision = Probabilité (d_j classé sous c_i = décision correcte) (1)

► Le *rappel* (degré de complétude) représente la probabilité conditionnelle que si un document d_j, pris au hasard, devrait être classifié sous la catégorie c_i, alors cette décision est prise [29]:

Rappel = Probabilité (d_j doit être classé sous c_i = décision prise) (2)

► La $F1$_measure. Définie par van Rijsbergen [30] elle sert à balancer les valeurs du rappel et de la précision. Sa forme générale introduit un paramètre β qui met en exergue la pondération différente du rappel (r) et de la précision (p) :

$$F\_Measure_{\beta}(r,p) = \frac{(\beta^2+1)pr}{\beta^2 p + r} \quad (3)$$

Afin d'attribuer une importance égale au *rappel* et à la *précision*, β est souvent remplacée par la valeur 1 [29] et F_1 prend alors la forme suivante :

$$F_1(r,p) = \frac{2pr}{p+r} \quad (4)$$

Ces mesures ne permettent pas de prendre en considération une liste de plusieurs classes pour la classification. Ces mesures ne sont pas adaptées à un classificateur multi-classes même si elles permettent d'avoir une idée initiale de la performance du classificateur.

3.5.2 Mesures visant les classifications multi-classes

Afin de comprendre le raisonnement du classificateur automatique et de l'exprimer, nous avons utilisé d'autres mesures découlant des premières (ci-dessus) qui mettent en évidence les performances de façon globale du système [31]. Ces mesures sont les suivantes :

► Pour un estimé moyen des performances du système de classification, on utilise la macro-moyenne qui attribue un même poids à toutes les catégories sans égard aux différences de taille entre les catégories.

Macro-Précision = somme (précisions)/nombre de classes (5)

► Pour évaluer les performances globales d'un système sans égard aux classes, on utilise la micro-moyenne qui attribue un poids égal à chaque document. Les catégories faiblement représentées auront plus d'influence selon leurs tailles.

Micro-Précision = somme des TP/(somme TP+Somme FP) (6)

3.6 Mesure hiérarchique

Comme les mesures de la classification plate et de la classification multi-classes ne respectent aucune des 3 conditions nécessaires dans le calcul des performances d'une classification hiérarchique multi-classes (voir les conditions à la section 3.2 de ce chapitre), nous avons utilisé la mesure de Kiritchenko qui est une mesure hiérarchique particulière hF exploitant le rappel et la précision non seulement au niveau de chaque classe mais surtout au niveau des ancêtres de chacun d'eux (étude réalisée par Kiritchenko & Al en 2005 [43]).

La mesure hF utilise les notions d'ancêtres pour calculer les erreurs de classification [44]. Formellement, en considérant une classification hiérarchique multi-étiquettes, on peut définir la mesure d'évaluation hF de la façon suivante [12]:

Pour toute instance (d_i,C_i) classifiée sous le sous-ensemble $C_i^{'}$ avec $C_i^{'} \subseteq C$, $d_i \in D$, $C_i \subseteq C$, on aura Les micro-moyennes hP et hR telles que :

$$hP = \frac{\sum_i \left| \text{Ancêtre}(C_i) \cap \textit{Ancêtre}(C_i^{'}) \right|}{\sum_i \left| \text{Ancêtre}(C_i^{'}) \right|}$$

$$hR = \frac{\sum_i \left| \text{Ancêtre}(C_i) \cap \textit{Ancêtre}(C_i^{'}) \right|}{\sum_i \left| \text{Ancêtre}(C_i) \right|}$$

La combinaison des deux valeurs hP et hR permet de calculer la F-Score (hF) :

$$hF_\beta = \frac{(\beta^2 + 1) hP.hR}{\beta^2 hP + hR},\ \beta \in [0,+\infty]$$

Afin de donner le même poids à la précision et au rappel, on utilise souvent $\beta = 1$ [29].

3.7 Mesure proposée

La mesure que nous proposons va se baser sur les valeurs récoltées d'un tableau de Sébastiani utilisé d'une façon différente.
Ainsi nous obtenons le tableau des valeurs qualitatives suivant :
Pour chaque document d_i classé par le classificateur multi-classes dans un ensemble de classes E_x, nous avons un autre ensemble de classes vraies E_{x1} choisi par l'expert du domaine :

3.7.1 Introduction

L'utilisation des méthodes classiques et élaborées citées ci haut se sont vite avérées insuffisantes pour la compréhension du raisonnement du classificateur comparé à celui des experts.
Une analyse des résultats de classification plus approfondie a été adoptée afin de comprendre la façon dont les nouvelles ont été classifiées. Cette méthode analyse les 3 types de listes utilisées (simple, hiérarchique et normée) et compare la classification des échantillons par ICM à celles effectuées par les 2 experts.

Voici la liste des possibilités issues de la comparaison entre l'expert et le classificateur :

1	**Compatibilité**	Le classificateur a choisi les mêmes classes choisies par l'expert. La colonne affiche le nombre de classes choisies.
2	**Enrichissement**	Le classificateur a choisi les mêmes classes choisies par l'Expert en plus d'autres classes que l'Expert n'a pas trouvé. La colonne affiche le nombre de classes trouvées en plus par le classificateur.
3	**Perte d'information**	Une classe au moins a été trouvée par le classificateur correspondant au choix de l'expert. La colonne affiche le nombre de classes trouvées par l'expert et non trouvées par le classificateur.
4	**Contradiction totale**	Aucune des classes choisies par l'expert n'a été trouvée par le classificateur. La colonne affiche le nombre de classes non trouvées.

Table 1: Liste des possibilités issues de la comparaison entre l'expert et le classificateur

Ces 4 groupes permettent de comprendre si une amélioration quelconque a été enregistrée au niveau de la classification en passant d'une liste simple à une liste hiérarchique, d'une liste simple à une liste normée et d'une liste hiérarchique à une liste normée.

3.7.2 Liste des 4 cas de figure généraux

1er cas, la compatibilité

La compatibilité a pour but initial de mesurer la stabilité du système de classification. Ainsi, en passant d'une liste à une autre, si le classificateur a choisi à chaque fois les mêmes classes par rapport à l'expert, cela signifie que le système est stable. Cette stabilité est très importante lorsqu'il s'agit de prendre des décisions éclairées dont la validité subsiste quelque soit l'ensemble de classes utilisé.
Dans le cas des autres mesures (voir ci-dessus) la stabilité n'est pas exprimée et donc n'est pas calculée ce qui enlève la possibilité de connaître le niveau de stabilité du système de classification utilisé.

2ème cas, l'enrichissement

L'enrichissement met l'accent sur les cas des classifications étendues. Ainsi, en utilisant les autres mesures de performance, les classes faisant partie d'un enrichissement étaient considérées comme mal choisies donc erronées. Dans ce cas précis, la mesure proposée va mettre l'accent sur ces cas particuliers qui ne devraient pas provoquer une baisse de performance mais plutôt une stabilité enrichie résultant d'un système de classification usant d'un bon apprentissage continu.

3ème cas, la perte d'information

La perte d'information représente les cas qui ne sont pas totalement faux ni totalement vrais de classification. Ainsi, si le classificateur retrouve au moins une des classes retrouvées par l'expert, cela signifie que l'erreur est moins grande que si toutes les classes trouvées étaient fausses. Le système est ainsi plus sensible aux cas partiellement vrais.

4ème cas, la contradiction totale

La contradiction totale représente la faillite du système de classification. En fait, moins on trouve de cas de contradiction totale, plus le système est fiable.

3.7.3 Qualité du classement

Après avoir dressé la liste des cas de figure possibles représentant le choix des classes par document dans une classification hiérarchique multi-classes, nous présentons ci bas la liste des possibilités détaillées selon chaque cas de figure lorsqu'on compare une liste simple à une liste hiérarchique ou normée et une liste hiérarchique à une liste normée.

1er cas, l'amélioration de la classification

Ce cas se présente lorsqu'au moins une classe vraie de plus a été trouvée par le classificateur. Ce cas de figure va augmenter la valeur de la performance si l'amélioration est totale et aussi lorsque le nombre de classes vraies trouvées en plus augmente (voir la Table 2).

2ème cas, l'enrichissement de la classification

L'enrichissement va pouvoir être mesuré selon le nombre de classes trouvées en trop par le classificateur. La stabilité représentée par la compatibilité est toujours plus avantageuse qu'un enrichissement proposant des classes dont le lien avec le document est discutable (voir la Table 3).

3ème cas, stabilité de la classification

Sachant que quelque soit les résultats d'une classification sur la base d'une liste simple versus une autre liste hiérarchique ou normée, la stabilité va toujours permettre de prouver que le le système est fiable. Ainsi, même si en utilisant la liste simple, les résultats ne sont pas probants, sachant que les autres listes fournissent les mêmes résultats, ceci est suffisant pour considérer la valeur stable du système de classification. Cette caractéristique est rarement prise en charge par les autres mesures. La meilleure des stabilités est un système compatible quelque soit la liste de classes utilisée (voir la Table 4).

4ème cas, diminution de la performance

Le cas le moins intéressant mais qui néanmoins permet de connaître le niveau de diminution de faible à important. Plus le nombre de classes vraies perdues en cours de la classification est grand, plus les performances du système vont diminuer. Le cas le plus grave est la perte totale de toutes les classes vraies (aucune n'a été trouvée dans la classification d'un document quelconque en passant d'une liste à une autre en comparant le classificateur à l'expert) (voir la Table 5).

Amélioration de la classification			
Code	**Signification**	**Condition (passage d'un état à un autre)**	**Explication additionnelle si nécessaire**
AP	**Amélioration partielle**	1. Contradiction totale - Perte d'information 2. Perte d'information - Perte d'information	Dans le 2ème cas, si le nombre de classes non trouvées a baissé.
AT	**Amélioration totale**	1. Perte d'information – Compatibilité 2. Perte d'information-Enrichissement 3. Contradiction totale – Compatibilité 4. Contradiction totale - Enrichissement	

Table 2: Liste des codes représentant l'amélioration des performances de la classification

Enrichissement			
Code	**Signification**	**Condition (passage d'un état à un autre)**	**Explication additionnelle si nécessaire**
ACE	**Augmentation de la compatibilité vers l'enrichissement**	Compatibilité - Enrichissement	
ES	**Enrichissement stable**	Enrichissement - Enrichissement	Si le nombre de classes trouvées en plus est le même dans les 2 listes
EA	**Enrichissement augmenté**	Enrichissement - Enrichissement	Si le nombre de classes trouvées en plus a augmenté.
ED	**Enrichissement diminué**	Enrichissement - Enrichissement	Si le nombre de classes trouvées en plus a diminué

Table 3: Liste des codes représentant l'enrichissement de la classification

Stabilité de la classification			
Code	**Signification**	**Condition (passage d'un état à un autre)**	**Explication additionnelle si nécessaire**
NBC	**Nouvelles bien classées**	Compatibilité - Compatibilité	
NBCPE	**Nouvelles bien classées avec perte d'enrichissement**	Enrichissement - Compatibilité	
NMC	**Nouvelles mal classées**	Contradiction totale - Contradiction totale	
NPBC	**Nouvelles partiellement bien classées**	Perte d'information - Perte d'information	Si le nombre de classes non trouvées est le même dans les 2 listes

Table 4: Liste des codes représentant la stabilité des performances de la classification

Diminution de la performance			
Code	**Signification**	**Condition (passage d'un état à un autre)**	**Explication additionnelle si nécessaire**
DP	**Diminution partielle**	1. Compatibilité - Perte d'information 2. Enrichissement - Perte d'information 3. Perte d'information - Perte d'information	Dans le 3ème cas, si le nombre de classes non trouvées a augmenté.
DT	**Diminution totale**	3 Compatibilité - Contradiction totale 4 Enrichissement - Contradiction totale 5 Perte d'information - Contradiction totale	

Table 5: Liste des codes représentant la diminution des performances de la classification

3.7.4 Mesure issue du calcul de la qualité de classement

Pour chaque document d_i classé par le classificateur dans l'ensemble E_x et par l'expert du domaine dans l'ensemble $E_{x'}$, nous avons la table de contingence par qualité de classement suivante :

Classes	Expert du domaine	Classificateur	Compatibilité (Stabilité)	Enrichissement	Perte d'information	Contradiction totale (Différence)
Classe$_{i0}$	ED_{i0}	C_{i0}				
Classe$_{i1}$	ED_{i1}	C_{i1}				
Classe$_{i2}$	ED_{i2}	C_{i2}	S_i	E_i	P_i	D_i
…	…	…				
Classe$_{in}$	ED_{in}	C_{in}				

Table 6: Qualité du classement par document

Le nombre de classes utilisées est n.
Les valeurs ED_{ij} et C_{ij} prennent 0 ou 1 selon le choix respectif de l'expert et du classificateur.
Les valeurs S_{ij}, E_{ij}, P_{ij} et D_{ij} sont calculées de la façon suivante :

$$S_i = \begin{cases} \sum_{j=1}^{n} ED_{ij} \text{ si } \forall j \in (1,..,n): ED_{ij} = C_{ij} \\ 0 \text{ sinon} \end{cases}$$

$$E_i = \begin{cases} \sum_{j=1}^{n} C_{ij} - \sum_{j=1}^{n} ED_{ij} \text{ si } \forall j \in (1,..,n): \left((ED_{ij} = 1) \Rightarrow (C_{ij} = 1)\right) \& \left(\exists j' \in (1,..,n): C_{ij'} = 1, D_{ij'} = 1\right) \\ 0 \text{ sinon} \end{cases}$$

$$P_i = \begin{cases} \sum_{j=1}^{n} ED_{ij} - \sum_{j=1}^{n} C_{ij} \text{ si } \left(\exists j \in (1,..,n): ED_{ij} = 1, C_{ij} = 1\right) \& \left(\exists j' \in (1,..,n): ED_{ij'} = 1, C_{ij'} = 0\right) \\ 0 \text{ sinon} \end{cases}$$

$$D_i = \begin{cases} \sum_{j=1}^{n} ED_{ij} \text{ si } \nexists j \in (1,..,n): ED_{ij} = 1 \Rightarrow C_{ij} = 1 \\ 0 \text{ sinon} \end{cases}$$

À partir des valeurs de qualité ci haut, on peut décrire pour chaque document l'amélioration ou la diminution de performance de classement en comparant la qualité du classement d'une liste par rapport à une autre de la façon suivante :

Pour un document d_i, on obtient le tableau qualitatif comparatif de passage d'une liste à une autre (voir les tables pour plus de détails sur les symboles utilisés : Table 2, Table 3, Table 4 et Table 5).
Le passage d'une liste à une autre prend 3 cas différents :
1er cas : L1 = Liste simple et L2 = Liste hiérarchique
2ème cas : L1 = Liste simple et L2 = Liste normée
3ème cas : L1 = Liste hiérarchique et L2 = Liste normée

Dans le tableau suivant, on considérera ce qui suit :

S_i : Classification stable du document d_i – Liste L1
S'_i : Classification stable du document d_i – Liste L2
E_i : Classification enrichie du document d_i – Liste L1
E'_i : Classification enrichie du document d_i – Liste L2
P_i : Classification avec perte d'information du document d_i – Liste L1
P'_i : Classification avec perte d'information du document d_i – Liste L2
D_i : Classification avec contradiction totale du document d_i – Liste L1
D'_i : Classification avec contradiction totale du document d_i – Liste L2

Différentes possibilités du passage de la liste L1 à la liste L2	**Qualité du passage**
$S_i \rightarrow S'_i$	**NBC**
$S_i \rightarrow E'_i$	**ACE**
$S_i \rightarrow P'_i$	**DP**
$S_i \rightarrow D'_i$	**DT**
$E_i \rightarrow S'_i$	**NBCPE**
$E_i \rightarrow E'_i$ Si $E_i(L1)=E'_i(L2)$	**ES**
$E_i \rightarrow E'_i$ Si $E_i(L1)<E'_i(L2)$	**EA**
$E_i \rightarrow E'_i$ Si $E_i(L1)>E'_i(L2)$	**ED**
$E_i \rightarrow P'_i$	**DP**
$E_i \rightarrow D'_i$	**DT**
$P_i \rightarrow S'_i$	**AT**
$P_i \rightarrow E'_i$	**AT**
$P_i \rightarrow P'_i$ Si $P_i(L1)=P'_i(L2)$	**NPBC**
$P_i \rightarrow P'_i$ Si $P_i(L1)<P'_i(L2)$	**DP**
$P_i \rightarrow P'_i$ Si $P_i(L1)>P'_i(L2)$	**AP**
$P_i \rightarrow D'_i$	**DT**
$D_i \rightarrow S'_i$	**AT**
$D_i \rightarrow E'_i$	**AT**
$D_i \rightarrow P'_i$	**AP**
$D_i \rightarrow D'_i$	**NMC**

Table 7: Détail de la qualité de la classification en comparant une liste de classes à une autre

En analysant les résultats du tableau ci haut, plusieurs possibilités se présentent lorsqu'une mesure spécifique est créée selon le besoin. On peut s'intéresser aux améliorations de performance et calculer ainsi les valeurs correspondantes telles les AT ou s'intéresser au côté enrichissant de la classification en mettant l'emphase sur les passage d'une classification mal

faite ou stable vers une classification apportant d'autres classes en plus de celles que l'expert a déjà désignées.
Cependant, pour cette étude, on est plus intéressés à prouver que le passage d'une liste simple vers une hiérarchique ensuite vers une normée va non seulement préserver la stabilité de la classification mais aussi l'enrichir et l'améliorer.
Dans ce but, si nous considérons que la stabilité ne peut influencer les performances d'un système de classification et que l'amélioration prouve son efficacité, nous avons calculé la fraction qui va permettre de comprendre le pourcentage des classifications améliorées par rapport au reste des classifications dont les performances ont eu un impact sur la qualité de la classification (performances améliorées et performances diminuées).

La fraction est représentée par la formule suivante :

Fraction = (Liste des classifications améliorées) / (Liste des classifications améliorées + Liste des classifications diminuées)

Avec les valeurs suivantes :

$$\textit{Liste des classifications améliorées} = \left(\sum_{d_i}^{i\in(1,..n)} AP + \sum_{d_i}^{i\in(1,..n)} AT \right)$$

$$\textit{Liste des classifications diminuées} = \left(\sum_{d_i}^{i\in(1,..n)} DP + \sum_{d_i}^{i\in(1,..n)} DT \right)$$

Ce qui donne la formule suivante représentant la mesure proposée:

$$\textit{Fraction} = \left(\sum_{d_i}^{i\in(1,..n)} AP + \sum_{d_i}^{i\in(1,..n)} AT \right) \Bigg/ \left[\left(\sum_{d_i}^{i\in(1,..n)} AP + \sum_{d_i}^{i\in(1,..n)} AT \right) + \left(\sum_{d_i}^{i\in(1,..n)} DP + \sum_{d_i}^{i\in(1,..n)} DT \right) \right]$$

Plus la fraction se rapproche du 100% plus la valeur de l'amélioration de la classification est importante.

Remarque : Dépendamment du résultat escompté et de la logique de classification utilisée, on peut aussi considérer comme une amélioration de la classification un enrichissement diminué, une compatibilité stable ou encore un passage d'un enrichissement vers une compatibilité. On considérera alors que la classification a diminué lorsqu'on a le cas d'un passage d'une compatibilité vers un enrichissement et aussi le cas d'un enrichissement augmenté. À partir de ce point de vue, on peut obtenir une mesure étendue de la façon suivante :

$$\textit{FractionÉtendue} = \left(\sum_{d_i}^{i\in(1,..n)} AP + \sum_{d_i}^{i\in(1,..n)} AT + \sum_{d_i}^{i\in(1,..n)} ED + \sum_{d_i}^{i\in(1,..n)} NBC + \sum_{d_i}^{i\in(1,..n)} NBCPE \right) \Bigg/$$

$$\left[\left(\sum_{d_i}^{i\in(1,..n)} AP + \sum_{d_i}^{i\in(1,..n)} AT + \sum_{d_i}^{i\in(1,..n)} ED + \sum_{d_i}^{i\in(1,..n)} NBC + \sum_{d_i}^{i\in(1,..n)} NBCPE \right) + \left(\sum_{d_i}^{i\in(1,..n)} DP + \sum_{d_i}^{i\in(1,..n)} DT + \sum_{d_i}^{i\in(1,..n)} ACE + \sum_{d_i}^{i\in(1,..n)} EA \right) \right]$$

3.7.5 Constat

Cette mesure proposée va au-delà de l'analyse quantitative de la classification puisqu'elle calcule avant tout la qualité de cette dernière. Elle prend notamment en considération des cas non pris en charge traditionnellement tels la compatibilité, l'enrichissement et le degré de diminution et d'augmentation de performance. Cette mesure est transférable est peut être utilisée pour calculer la qualité du classement de n'importe quel autre système de classification basé sur n'importe quel domaine et utilisant n'importe quel ensemble de classes.

CHAPITRE 4

Méthodologie

4.1 Introduction

Dans ce quatrième chapitre, nous allons présenter la procédure adoptée pour la préparation de l'environnement de simulation, le calcul des résultats de classification et l'analyse de ces derniers.

Le passage d'une base de données brute Reuters RCV1 vers un système de simulation efficace a nécessité un ensemble d'étapes importantes permettant d'entraîner le classificateur, de classifier des échantillons de nouvelles et enfin d'analyser des résultats de classification.
Dans ce qui suit, nous présentons la liste des étapes importantes suivies.

Tel que montré à la Figure 2, notre méthodologie itérative comporte quatre étapes séquentielles (voir l'annexe H pour connaître le cycle de vie du projet du mémoire) :

1. Le développement d'une liste de sujets (mots clés) et des ontologies pour la classification
2. L'échantillonnage des nouvelles sur lesquelles les sujets seront appliqués par des experts
3. L'évaluation du classificateur sur les échantillons d'entraînement et de test
4. La comparaison et l'interprétation des résultats des divers tests

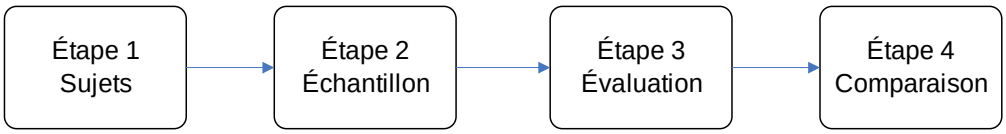


Figure 2: Procédure itérative

4.2 Extraction, épuration et échantillonnage des données de simulation

Les données de simulation sont issues du corpus RCV1 de Reuters (voir Annexe I) qui tournent autour d'un ensemble riche de nouvelles financières.

Afin d'obtenir une base de données relationnelle opérationnelle pour la simulation, les étapes suivantes ont été nécessaires :

1. Élimination du bruitage qui concerne les nouvelles redondantes et les codes inutilisés (voir la section I.2 de l'annexe I). Cette étape a permis de créer la version2 de Reuters RCV1.
2. Extraction de 41 214 nouvelles liées au domaine de la fusion et acquisition (*Mergers and Acquisitions*) (voir les sections I.1 et I.4 de l'annexe I).
3. Extraction aléatoire d'un échantillon de 1000 nouvelles pour les besoins de la simulation. (voir la section I.5 de l'annexe I).

4.3 Développement des listes de classes

À cette étape, une liste de sujets (mots clés) pertinents au domaine des *fusions* et *acquisitions* est désignée en se limitant aux spécificités des 1000 nouvelles de l'échantillon.

Comme le but est de comparer la classification normée à celles simples et hiérarchisées limitées, cette procédure sera alors exécutée en trois itérations. Ainsi, selon le type de classification à considérer, nous aurons :

1. **Simple** : Les sujets seront une simple liste de sujets pertinents au domaine des *fusions* et *acquisitions*. Nous identifions une liste de sujets standards présents dans les modèles issus de la littérature. Ces sujets permettent de déterminer si une nouvelle traite de la structure, des raisons, ou des résultats d'une transaction.

    1. ***Merger***
    2. ***Acquisition***
    3. ***Price***
    4. ***Control***
    5. ***Integration***
    6. ***Value***
    7. ***Revenues***
    8. ***Profits***
    9. ***Assets***
    10. ***Debt***

2. **Hiérarchique** : Les sujets seront une hiérarchie limitée à deux niveaux de sujets pertinents au domaine des *fusions* et *acquisitions*. Nous tentons d'identifier au niveau supérieur les 3 types de nouvelles, soit raisons, structure, et résultats. Nous organisons ensuite les concepts de la liste simple sous chacun des 3 sujets, selon leur fréquence de mention dans la littérature.

    1. ***Structure***
        - 1.1. ***Merger***
        - 1.2. ***Acquisition***
        - 1.3. ***Price***
        - 1.4. ***Control***
    2. ***Reasons***
        - 2.1. ***Integration***
        - 2.2. ***Value***
    3. ***Results***
        - 3.1. ***Revenues***
        - 3.2. ***Profits***
        - 3.3. ***Assets***
        - 3.4. ***Debt***

3. **Normée :** Les sujets seront une hiérarchie riche à plusieurs niveaux de sujets pertinents au domaine des *fusions* et *acquisitions*, tous pris à partir du schéma XSD (*XML Schema Definition*) de la norme XBRL.

1. ***Income***
    - 1.1. ***Revenues***
        - 1.1.1. ***Sales***
        - 1.1.2. ***Costs***
        - 1.1.3. ***Gross***
    - 1.2. ***Expenses***
        - 1.2.1. ***Distribution***
        - 1.2.2. ***Administrative***
        - 1.2.3. ***Other***
    - 1.3. ***Profits***
        - 1.3.1. ***Interests***
        - 1.3.2. ***Taxes***
        - 1.3.3. ***Depreciation***
        - 1.3.4. ***Impairment***
        - 1.3.5. ***Earnings***
2. ***Financial***
    - 2.1. ***Assets***
        - 2.1.1. ***Property***
        - 2.1.2. ***Investment***
        - 2.1.3. ***Goodwill***
        - 2.1.4. ***Inventory***
        - 2.1.5. ***Receivables***
        - 2.1.6. ***Cash***
    - 2.2. ***Liabilities***
        - 2.2.1. ***Debt***
        - 2.2.2. ***Equity***
3. ***Notes***
    - 3.1. ***Reasons***
        - 3.1.1. ***Integration***
        - 3.1.2. ***Value***
    - 3.2. ***Structure***
        - 3.2.1. ***Merger***
        - 3.2.2. ***Acquisition***
        - 3.2.3. ***Price***
        - 3.2.4. ***Control***

Les sujets pour les 3 types de classifications ont été sélectionnés sur la base de la littérature récente dans le champ des sciences administratives, plus précisément en gestion stratégique et en finance, qui sont les 2 disciplines les plus actives dans l'étude des *fusions* et des *acquisitions* [20]. Pour le choix des sujets, nous nous concentrons sur les variables comptables qui expliquent généralement la majorité du comportement et des décisions des firmes acquéreurs, ainsi que les interprétations faites par les analystes dans les nouvelles financières [25, 26].

Nous cherchons ici à classifier les nouvelles selon qu'elles discutent des sujets liés à la « structure » de la transaction, aux « raisons » justifiant la transaction, et aux « résultats » de la transaction. La structure de la transaction est particulièrement affectée par les options possibles pour la combinaison des firmes en jeu [21, 22] :

1. Les raisons sont souvent liées à des phénomènes circonstanciels de l'industrie en question, telles que l'incertitude, la croissance, la concurrence, ou les synergies entre firmes d'un même secteur [23-25].
2. Les résultats d'une transaction sont associés de près à la performance boursière, à la vitesse d'intégration des firmes, et à l'impact sur la création de valeur (exemple : innovation) de la nouvelle entité [26-28].

4.4 Classification manuelle experte

Une fois l'échantillonnage des 1000 nouvelles finalisé, nous avons procédé à la classification manuelle des données ciblées avec l'aide de deux experts du domaine.

Les deux experts travaillent indépendamment l'un de l'autre et classifient chacun une copie de l'échantillon sur la base des sujets hiérarchiques normés choisis.

Une application développée sous Access est utilisée afin de faciliter l'analyse de chacune des nouvelles de l'échantillon, leur classification, et leur récupération en vue du prochain processus.
Pour plus de détails sur les étapes suivies par les experts pour la classification manuelle, se référer à l'annexe J.

Note : Un troisième expert du domaine est désigné pour son expertise professionnelle et étendue en finances. Ce dernier se charge de corriger les anomalies de classification des nouvelles en diminution totale de performance après l'analyse des résultats de la classification automatique à travers la mesure proposée.

4.5 Convergence des experts

4.5.1 Introduction

Étant donné que deux classifications manuelles différentes ont été réalisées sur la base d'un même échantillon de 1000 nouvelles financières en *fusion*/*acquisition*, une validation des résultats est nécessaire afin de vérifier le nombre de fausses classifications (documents classés dans des classes erronées) et le niveau d'erreur observé selon la distance des classes vraie et fausse (voir les 3 conditions au niveau de la section 3.4). Le résultat de cette validation permettra de développer une classification manuelle correcte à utiliser dans l'évaluation du classificateur commercial ICM.

4.5.2 Raffinement du choix des classes et des nouvelles

Afin de faire une analyse riche de la dispersion et du poids de chaque classe et de chaque nouvelle, nous nous sommes appuyés sur l'expertise récupérée de la classification manuelle précédente. C'est ainsi que, grâce à une étude plus ciblée et plus directe de la classification des experts nous avons pu dresser une liste des caractéristiques formant les prises de décision des 2 experts selon ce qui suit (voir la Table 8 et la Table 11):

- Certaines classes sont dominantes telles les classes ***Acquisition***, ***Sales*** et ***Merger***.
- D'autres classes sont peu fréquemment référées par les 2 experts. Nous pouvons citer les classes: ***Gross***, ***Depreciation***, ***Other***, ***Impairment***, ***Inventory***…
- En comparant la classification de chaque expert, nous pouvons déduire les nouvelles qui sont classées de façon compatible, les nouvelles qui ont connu un enrichissement, une perte d'information et une contradiction totale (voir la **Erreur ! Source du renvoi introuvable.**). Ainsi, le nombre de nouvelles classées de façon similaire par les 2 experts est égal à 81, ce qui permet de favoriser leur utilisation dans l'entraînement du classificateur (voir la Table 9).

Cependant, étant donné que le nombre de nouvelles compatibles est insuffisant pour entraîner le classificateur, il est judicieux d'y intégrer des nouvelles partiellement compatibles dans lesquelles les classes dominantes sont présentes. Ces dernières seront basées sur l'enrichissement et la perte d'information. Ainsi, plus ces 2 dernières valeurs sont basses, plus la différence de classification est moins importante; Exemple : une nouvelle qui a été classée par l'expert 1 dans 4 classes différentes de celles de l'expert 2, est moins intéressante qu'une nouvelle classée dans 2 classes différentes de l'expert 2 (voir la Table 10). En regardant la Table 9, nous remarquons que le nombre de nouvelles affectées par l'enrichissement et la compatibilité est important. En éliminant les nouvelles contradictoires par rapport aux experts (classifications totalement différentes), nous obtenons un nombre de nouvelles utilisable pour l'étude équivalent à 779 nouvelles (car il y a 221 nouvelles contradictoires) parmi lesquelles se trouvent les 81 nouvelles compatibles.

Nombre de classes utilisées par chaque expert (Recherche des classes dominantes)		
Classes	**Expert 1**	**Expert 2**
Sales	297	341
Costs	39	142
Gross	3	18
Distribution	13	82
Administrative	158	79
Other	15	1
Interests	14	71
Taxes	19	50
Depreciation	2	70
Impairment	13	5
Earnings	108	88
Property	128	165
Investment	83	109
Goodwill	16	33
Inventory	30	8
Receivables	26	119
Cash	103	133
Debt	72	62
Equity	93	47
Integration	21	111
Value	59	133
Merger	224	244
Acquisition	537	660
Price	296	113
Control	105	54

Table 8: Nombre de classes utilisées par chaque expert

	Nombre de nouvelles	
	Exp1-Exp2	***Exp2-Exp1***
Compatibilité	81	81
Enrichissement	99	231
Perte d'information	604	530
Contradiction totale	216	158

Table 9: Comparaison de la classification des nouvelles par les 2 experts.

	Enrichissement (nombre de classes en plus)						
	1	**2**	**3**	**4**	**5**	**6**	**>6**
Exp1-Exp2	58	32	6	3	0	0	0
Exp2-Exp1	118	64	27	16	5	0	1

	Perte d'information (nombre de classes en moins)						
	1	**2**	**3**	**4**	**5**	**6**	**>6**
Exp1-Exp2	256	191	87	48	16	4	2
Exp2-Exp1	263	168	67	20	7	4	1

Table 10: Nombre de nouvelles affectée par un enrichissement ou par une perte d'information

4.5.3 Choix des classes à utiliser avec ICM

Quant nous comparons l'expert 1 à l'expert 2, nous remarquons que certaines classes ont été privilégiées par les 2 (voir la Table 11). Ce qui permet de définir certaines contraintes sur l'arbre des classes. En considérant le nombre de nouvelles classifiées partiellement ou totalement de la même façon par les 2 experts, on note que certaines classes sont plus privilégiées que d'autres (exemple : La classe *Acquisition* qui a été choisie 447 fois de la même façon par les 2 experts). D'autres classes sont d'influence négligeable (exemple : la classe *Inventory* qui n'a pas été choisie en même temps par les 2 experts).

Plus l'influence d'une classe est minimale, plus son impact sur la classification est négatif et donc peut induire en erreur (car il n'y a pas d'accord entre les 2 experts dans la classification). Afin d'obtenir un nombre satisfaisant de nouvelles pour l'entraînement d'ICM et pour les tests de classification, nous avons utilisé la méthode des sommes qui permet de descendre le seuil d'utilisation similaire jusqu'à ce que le nombre total de nouvelles à extraire est satisfaisant. Pour cela, en fixant le seuil à 17 (nombre de fois minimal que la classe a été simultanément choisie par les 2 experts), alors les nouvelles à extraire seront celles que les experts auront classifié de la même façon dans une ou plusieurs des 14 classes ci bas (voir la Table 11).

Rang de la classe selon son importance	Classes	Nombre de fois utilisées de façon similaire par les 2 experts
1	**Acquisition**	447
2	**Merger**	198
3	**Sales**	126
4	**Cash**	94
5	**Price**	62
6	**Earnings**	52
7	**Administrative**	49
8	**Investment**	40
9	**Debt**	38
10	**Costs**	30
11	**Property**	26
12	**Value**	24
13	**Taxes**	17
14	**Control**	17
15	**Equity**	10
16	**Distribution**	8
17	**Goodwill**	6
18	**Integration**	6
19	**Interests**	5
20	**Receivables**	5
21	**Gross**	2
22	**Depreciation**	2
23	**Other**	0
24	**Impairment**	0
25	**Inventory**	0

Table 11: Nombre de fois que chaque classe a été choisie par les 2 experts en même temps

4.6 Entraînement du classificateur et choix des échantillons

4.6.1 Choix des classes pour l'entraînement d'ICM

En se basant sur la logique de la convergence des 2 experts selon la section précédente, la liste des classes dominantes suivantes a été sélectionnée pour l'entraînement et la simulation à travers ICM (voir la Table 12):

Classes dominantes	Nombre de nouvelles choisies pour l'entraînement
Taxes	6
Control	6
Value	8
Property	9
Costs	10
Debt	12
Investment	13
Administrative	16
Earnings	17
Price	20
Cash	31
Sales	42
Merger	66
Acquisition	149

Table 12: Nombre de nouvelles choisies pour le pré entraînement et pour l'échantillon

Ce choix de classes a permis de déduire les listes de sujets à utiliser dans la simulation :

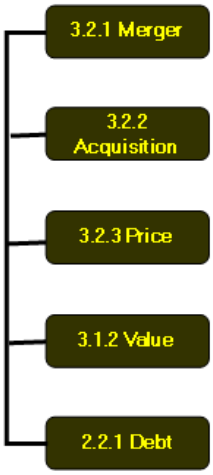


Figure 3: Liste simple

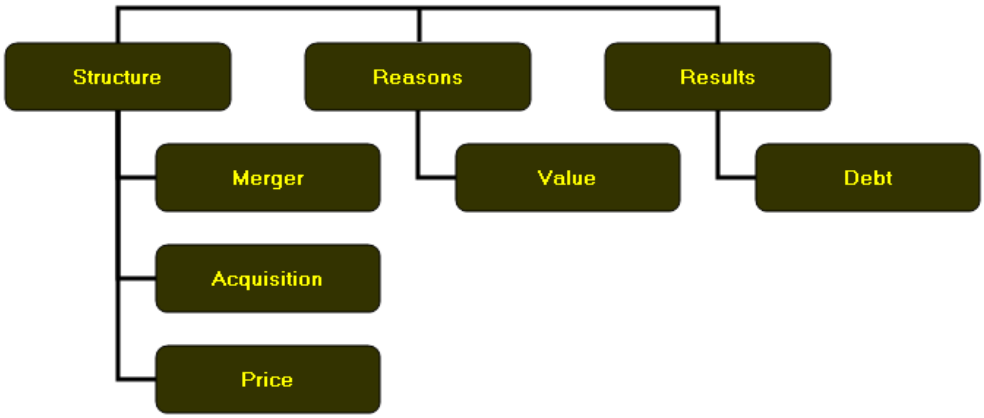


Figure 4: Liste hiérarchique

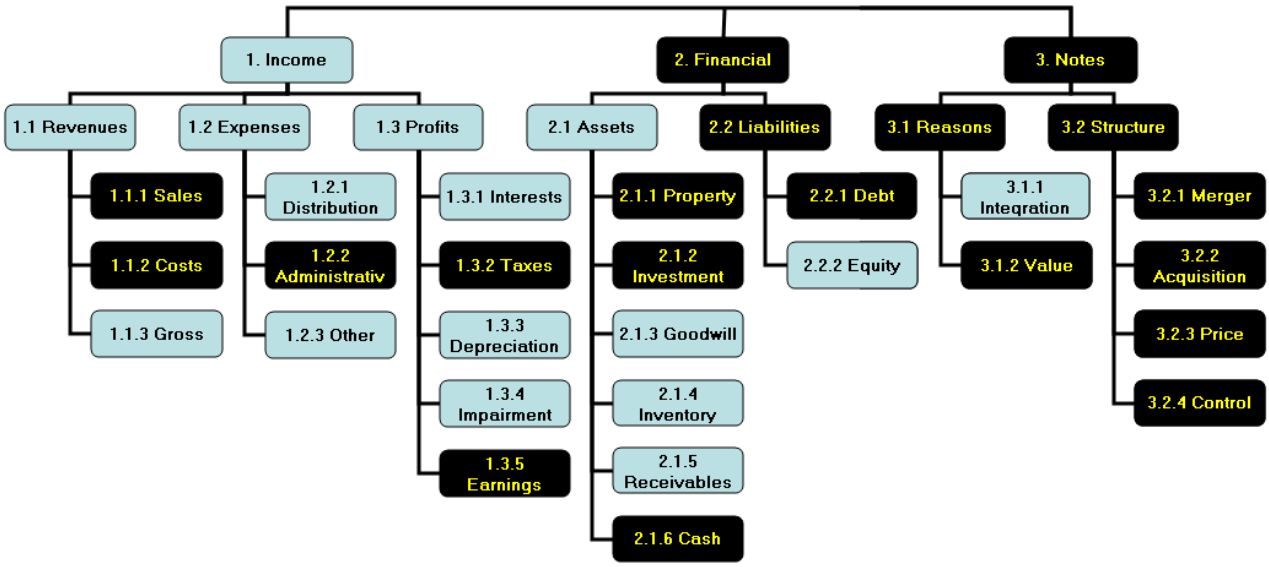


Figure 5: Liste normée réduite (classes sous fond noir)

4.6.2 Choix des nouvelles pour l'entraînement d'ICM et pour la simulation

Avant tout, en se basant sur les nouvelles classées de façon compatibles et celles classées de la même façon mais avec un enrichissement en plus, nous avons choisi un ensemble d'entraînement de 201 nouvelles basé sur l'utilisation du tiers des classes dominantes (voir la Table 12) et 201 nouvelles pour l'échantillon en utilisant un autre tiers des classes dominantes.

Il existe 6 ensembles d'entraînement (de format XML – voir la Figure 6) pour chacune des listes et pour chaque expert (voir la Table 13).

	Expert 1	Expert 2
Liste simple	ClassesDominantesEnsemblePreentrainementListeSimpleE1.xml	ClassesDominantesEnsemblePreentrainementListeSimpleE2.xml
Liste hiérarchique	ClassesDominantesEnsemblePreentrainementListeHierarE1.xml	ClassesDominantesEnsemblePreentrainementListeHierarE2.xml
Liste normée	ClassesDominantesEnsemblePreentrainementListeNormeeE1.xml	ClassesDominantesEnsemblePreentrainementListeNormeeE2.xml

Table 13: Ensembles d'entraînement utilisés dans la création des bases de connaissance

Soit un exemple d'un document XML d'entraînement :

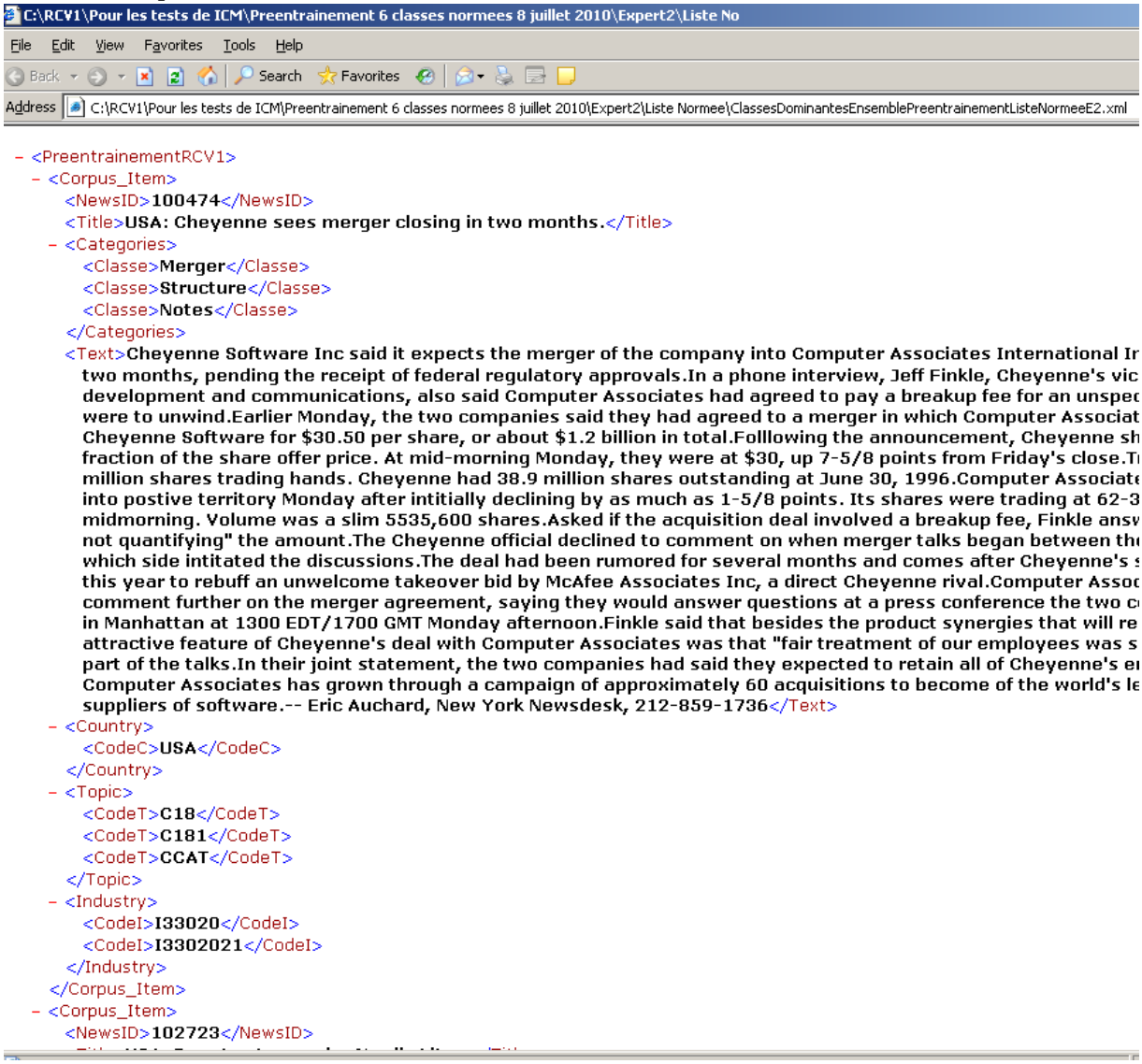


Figure 6: Exemple d'un document XML servant à l'entraînement du classificateur ICM

On constatera qu'une nouvelle faisant partie d'un ensemble d'entraînement, va contenir les champs suivant :

Champs contenant de l'information :
NewsID : Contenant le code de la nouvelle
Title : Titre de la nouvelle

Champs servant à la classification (contenant des classes)
Classe : Va contenir un des sujets contenus dans la liste de classes utilisée (Liste simple, hiérarchique ou normée)

Champs contenant de l'information NLP utilisée dans le raisonnement du classificateur :
Text : Contient le corps de la nouvelle. Cette information va être analysée par le moteur de conception des relations RME d'ICM (voir Annexe A).

Champs hérités de Reuters RCV1 dont l'information est secondaire :
CodeC : Code *Country*
CodeI : Code *Industry*
CodeT : Code *Topic*

Ces informations sont représentées par ICM selon la figure suivante :

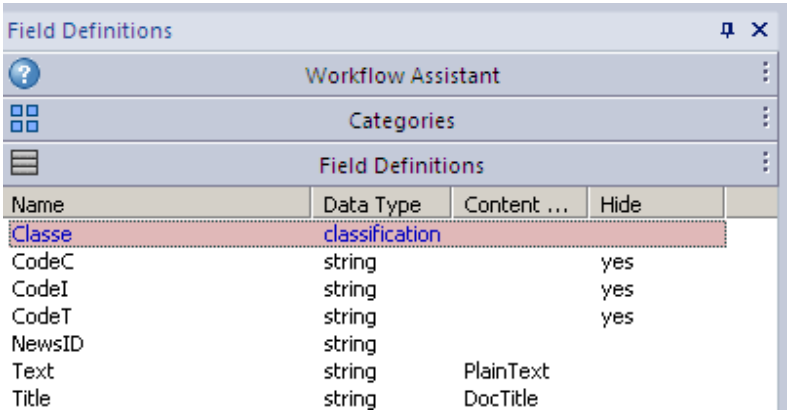


Figure 7: Spécification des champs contenant les classes et des champs NLP (Texte)

4.6.3 Entraînement d'IBM Classification Module

La classification des échantillons de nouvelles choisis précédemment va être effectuée au sein du Workbench qui est un module d'ICM.
Les 3 étapes suivantes sont nécessaires afin d'entraîner ICM (pour plus de détails sur le fonctionnement d'ICM, se reporter au Annexe A) :

1. Création et calibrage de bases de connaissance sur la base de contenus pré catégorisés (ensembles d'entraînement).
2. Pré - entraînement du classificateur
3. Création des plans de décision sur la base des mêmes contenus pré catégorisés ci dessus.

4.6.3.1 Création et calibrage des bases de connaissance

ICM devant être entraîné avant de pouvoir le lancer sur une classification automatique, il est nécessaire d'utiliser des bases de connaissance basées sur les 3 listes d'entraînement (liste simple, liste hiérarchique et liste normée) citées à la section 4.6.2.
Les étapes suivantes vont permettre de configurer les bases de connaissance :

Étape 1 : Création des bases de connaissance.
Pour chaque expert, nous avons créé 3 types de bases de connaissance qui seront chacune basée sur l'une des 3 listes (simple, hiérarchique et normée). Nous obtiendrons ainsi 6 bases de connaissance que nous pouvons voir dans la Table 14. Ces bases de connaissance ont un format reconnu d'ICM et peuvent être réutilisées pour d'autres classifications d'échantillons de nouvelles tournant autour des mêmes sujets (classes).

	Expert 1	Expert 2
Liste simple	E1-6Classes-ListeSimpleKB	E2-6Classes-ListeSimpleKB
Liste hiérarchique	E1-6Classes-ListeHierarchiqueKB	E2-6Classes-ListeHierarchiqueKB
Liste normée	E1-6Classes-ListeNormeeKB	E2-6Classes-ListeNormeeKB

Table 14: Liste des bases de connaissance utilisées dans l'entraînement d'ICM selon le cas

Étape 2 : Configuration des bases de connaissance.
La configuration des bases de données représentant le raisonnement de chaque expert sur la base des listes de classes correspondantes se fait en identifiant la liste des champs utiles pour le raisonnement d'ICM et pour la classification.
Lors de cette configuration, les champs représentant des classes et ceux représentant un contenu NLP permettant de déduire le raisonnement sont spécifiés (Figure 7).
La structure des bases de connaissance est déterminée à travers la mise en place de l'arborescence des classes. Ainsi, pour les bases de connaissance basées sur la liste normée, la structure sous ICM est spécifiée de la façon suivante (Figure 8):

Figure 8: Base de connaissance basée sur la liste normée de classes

4.6.3.2 Pré entraînement du classificateur ICM

L'initiation de l'apprentissage utilise les bases de connaissance créées.
À cette étape, nous spécifions à ICM que la base de connaissance créée servira pour l'apprentissage et sera utilisée dans les prochaines classifications automatiques de documents.

Il est important de garder l'environnement initial d'ICM afin de classer les nouvelles selon le même principe et le même environnement car les résultats de la classification ne doivent pas être influencés par un apprentissage continu qui risque de corrompre d'une façon ou d'une autre les performances du classificateur (voir la Figure 9 représentant l'apprentissage choisi lors de la création de la Base de connaissance).

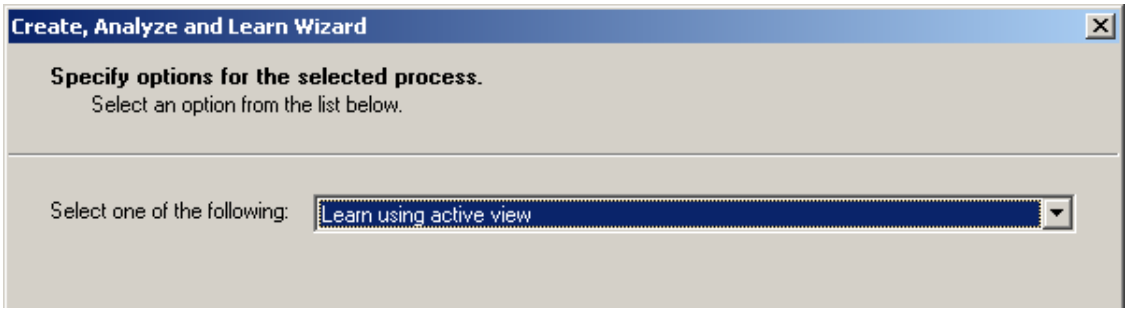


Figure 9: Préparation de la base de connaissance pour l'apprentissage

4.6.3.3 Création des plans de décision

La création des bases de connaissance précédentes ainsi que le pré entraînement d'ICM ne sont pas suffisants pour préparer le terrain pour une classification automatique des échantillons des nouvelles de Reuters. Pour cela, il faudra avant tout créer une collection de règles afin de

spécifier à ICM la façon dont les documents seront classifiés. Ce qui implique la nécessité de créer des plans de décision pour chacune des bases de connaissance.
La création d'un plan de décision se fait de la même façon que celle d'une base de connaissance et se base sur le même contenu puisqu'une analyse des données est nécessaire afin de vérifier la validité des règles créées.

Voici les étapes suivies dans la création de chacun des plans de décision :
1. Création des projets de plan de décision.
2. Ajout d'une base de connaissance correspondante à un projet particulier.
3. Définition des règles de décision.
4. Analyse du plan de décision par rapport aux données d'entraînement actuelles.

Étape 1 : Création des projets de plan de décision.
La création d'un projet de plan de décision se fait en sélectionnant un contenu pour chacun des cas des listes de classes : simple, hiérarchique et normé.
Ce contenu est le même que celui utilisé lors de la création de la base de connaissance correspondante.

Étape 2 : Ajout d'une base de connaissance correspondante à un projet particulier.
Il est important, avant de spécifier les règles et les déclencheurs de règles (*triggers*), d'ajouter la base de connaissance correspondante au projet (Figure 10). Cette liaison permet aux règles de se référer aux champs de la base de connaissance citée (voir la Table 15).

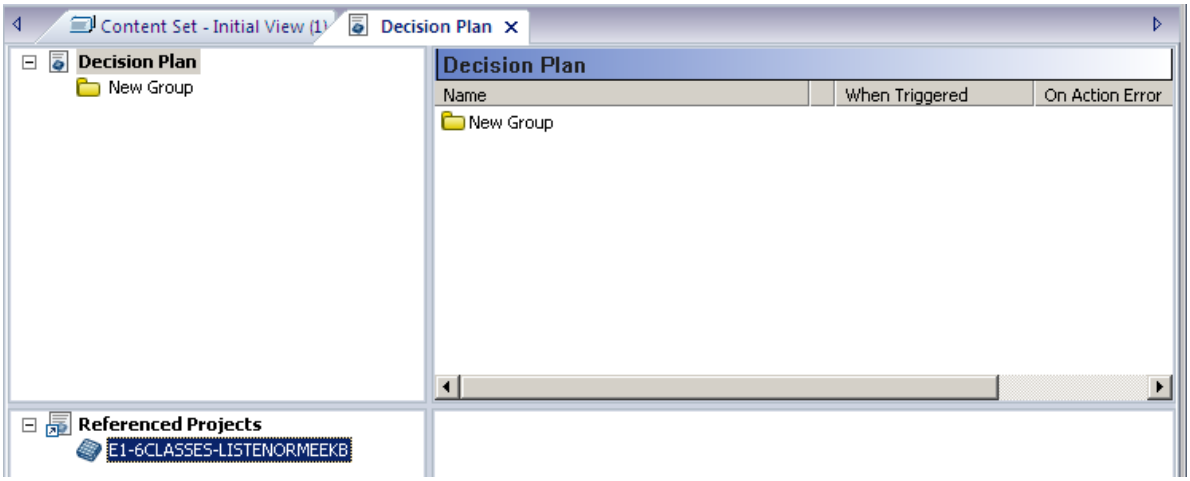


Figure 10: Liaison entre un plan de décision et la structure d'une base de connaissance

Listes correspondantes		*Projets de plan de décision*	*Bases de connaissance correspondantes*
Liste simple	**Expert 1**	E1-6Classes-ListeSimpleDP	E1-6Classes-ListeSimpleKB
	Expert 2	E2-6Classes-ListeSimpleDP	E2-6Classes-ListeSimpleKB
Liste hiérarchique	**Expert 1**	E1-6Classes-ListeHierarchiqueDP	E1-6Classes-ListeHierarchiqueKB
	Expert 2	E2-6Classes-ListeHierarchiqueDP	E2-6Classes-ListeHierarchiqueKB
Liste normée	**Expert 1**	E1-6Classes-ListeNormeeDP	E1-6Classes-ListeNormeeKB
	Expert 2	E2-6Classes-ListeNormeeDP	E2-6Classes-ListeNormeeKB

Table 15: Liste des plans de décision et des bases de connaissance correspondantes

Étape 3 : Définition des règles de décision.
Les règles de décision vont permettre d'affecter des classes à des nouvelles selon des critères de raisonnement précis qui tournent autour du niveau de correspondance selon un score donné entre un document et ses classes (voir l'Annexe A pour plus de détails).
Le score est calculé selon les niveaux de seuil de représentativité des classes dans la base de connaissance. Une règle de décision servira en outre à spécifier un seuil minimal pour choisir une ou des classes représentant la nouvelle courante (Figure 11).

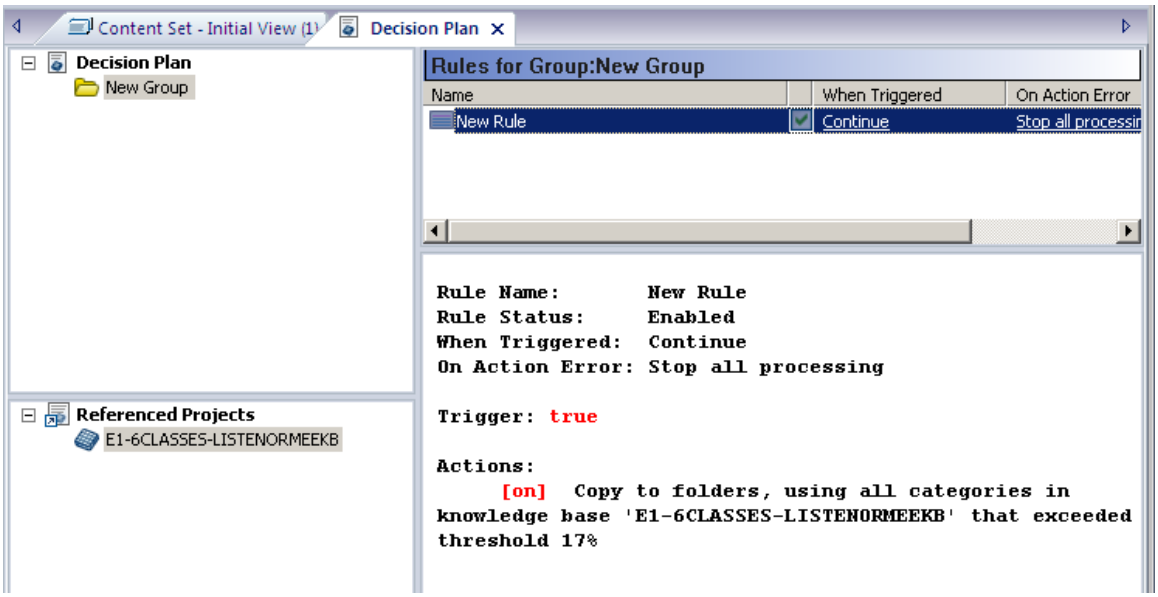
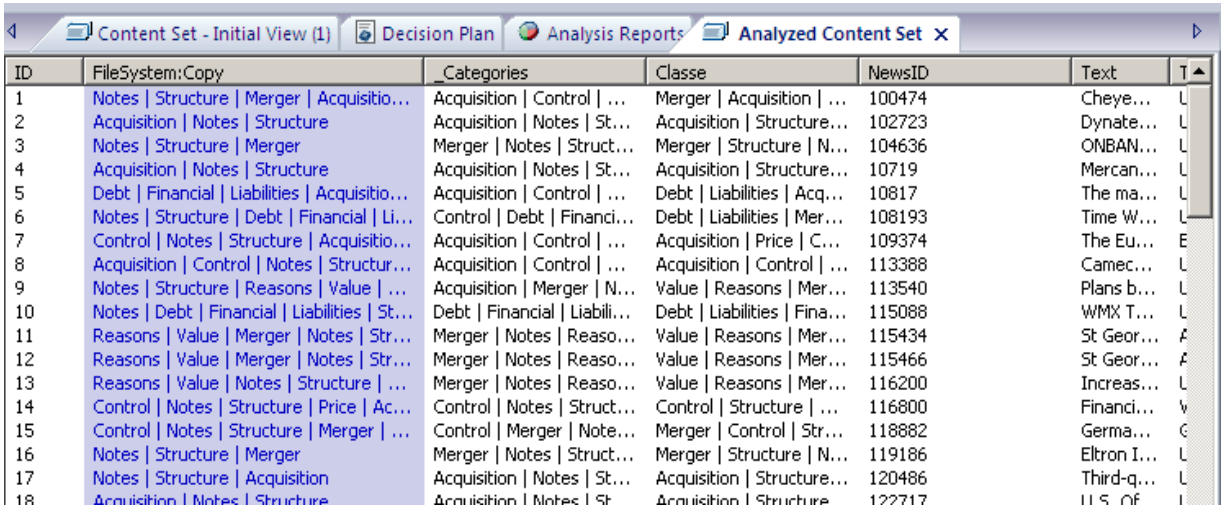


Figure 11: Exemple de règle décision

Étape 4 : Analyse du plan de décision par rapport aux données d'entraînement actuelles.
L'analyse du plan de décision permet de vérifier si les règles créées permettent une classification rapprochée du raisonnement de l'entraînement et réduisent le pourcentage d'erreur de classification en englobant la plupart des classes spécifiées par l'expert dans un choix fait par la classification automatique d'ICM (Figure 12).



ID	FileSystem:Copy	_Categories	Classe	NewsID	Text
1	Notes | Structure | Merger | Acquisitio...	Acquisition | Control | ...	Merger | Acquisition | ...	100474	Cheye...
2	Acquisition | Notes | Structure	Acquisition | Notes | St...	Acquisition | Structure...	102723	Dynate...
3	Notes | Structure | Merger	Merger | Notes | Struct...	Merger | Structure | N...	104636	ONBAN...
4	Acquisition | Notes | Structure	Acquisition | Notes | St...	Acquisition | Structure...	10719	Mercan...
5	Debt | Financial | Liabilities | Acquisitio...	Acquisition | Control | ...	Debt | Liabilities | Acq...	10817	The ma...
6	Notes | Structure | Debt | Financial | Li...	Control | Debt | Financi...	Debt | Liabilities | Mer...	108193	Time W...
7	Control | Notes | Structure | Acquisitio...	Acquisition | Control | ...	Acquisition | Price | C...	109374	The Eu...
8	Acquisition | Control | Notes | Structur...	Acquisition | Control | ...	Acquisition | Control | ...	113388	Camec...
9	Notes | Structure | Reasons | Value | ...	Acquisition | Merger | N...	Value | Reasons | Mer...	113540	Plans b...
10	Notes | Debt | Financial | Liabilities | St...	Debt | Financial | Liabili...	Debt | Liabilities | Fina...	115088	WMX T...
11	Reasons | Value | Merger | Notes | Str...	Merger | Notes | Reaso...	Value | Reasons | Mer...	115434	St Geor...
12	Reasons | Value | Merger | Notes | Str...	Merger | Notes | Reaso...	Value | Reasons | Mer...	115466	St Geor...
13	Reasons | Value | Notes | Structure | ...	Merger | Notes | Reaso...	Value | Reasons | Mer...	116200	Increas...
14	Control | Notes | Structure | Price | Ac...	Control | Notes | Struct...	Control | Structure | ...	116800	Financi...
15	Control | Notes | Structure | Merger | ...	Control | Merger | Note...	Merger | Control | Str...	118882	Germa...
16	Notes | Structure | Merger	Merger | Notes | Struct...	Merger | Structure | N...	119186	Eltron I...
17	Notes | Structure | Acquisition	Acquisition | Notes | St...	Acquisition | Structure...	120486	Third-q...
18	Acquisition | Notes | Structure	Acquisition | Notes | St...	Acquisition | Structure...	122717	U.S. Of...

Figure 12: Résultats de l'analyse d'un projet de plan de décision

4.7 Classification automatique

La classification automatique de l'échantillon de nouvelles va s'appuyer sur le raisonnement du classificateur entraîné.

Une fois les bases de connaissance et les projets de plans de décision créés et configurés, il est plus aisé de lancer la classification automatique d'ICM sur les échantillons de nouvelles précédemment sélectionnés. Pour cela, il faut respecter un ensemble d'étapes qui permettront de récupérer des listes de nouvelles classifiées sur lesquelles des calculs de performance seront effectués afin de vérifier si les buts spécifiés au début de ce projet ont été atteints.

4.7.1 Démarrage des projets de plans de décision et des bases de connaissance

Cette étape permet de démarrer un plan de projet spécifique à une liste simple, hiérarchique ou normée ainsi que de la base de connaissances correspondante grâce à la console de gestion du module de classification d'ICM (Figure 13).

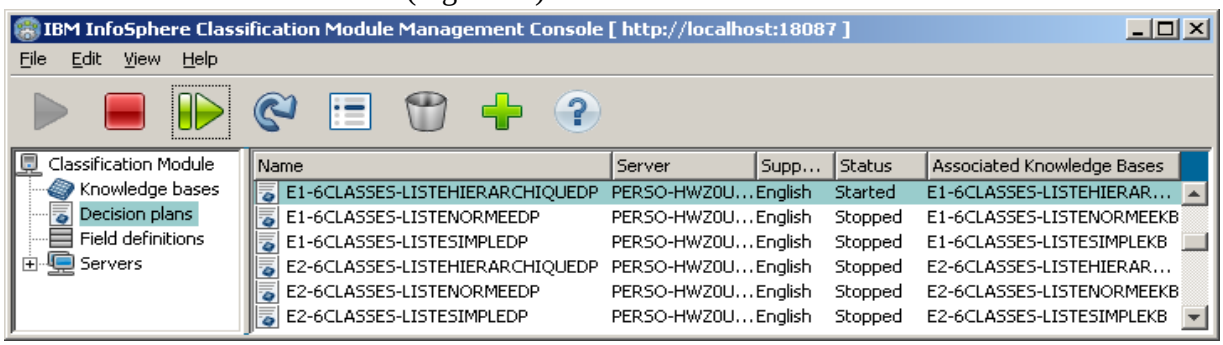


Figure 13: Console de gestion du module de classification automatique d'ICM.

4.7.2 Classification des échantillons de nouvelles

La classification de documents sous ICM se fait à l'intérieur du centre de classification qui permet de choisir le plan de décision à exécuter, le contenu à classer (échantillon de nouvelles sous format XML) et enfin l'endroit dans lequel les résultats seront disposés (Figure 14).

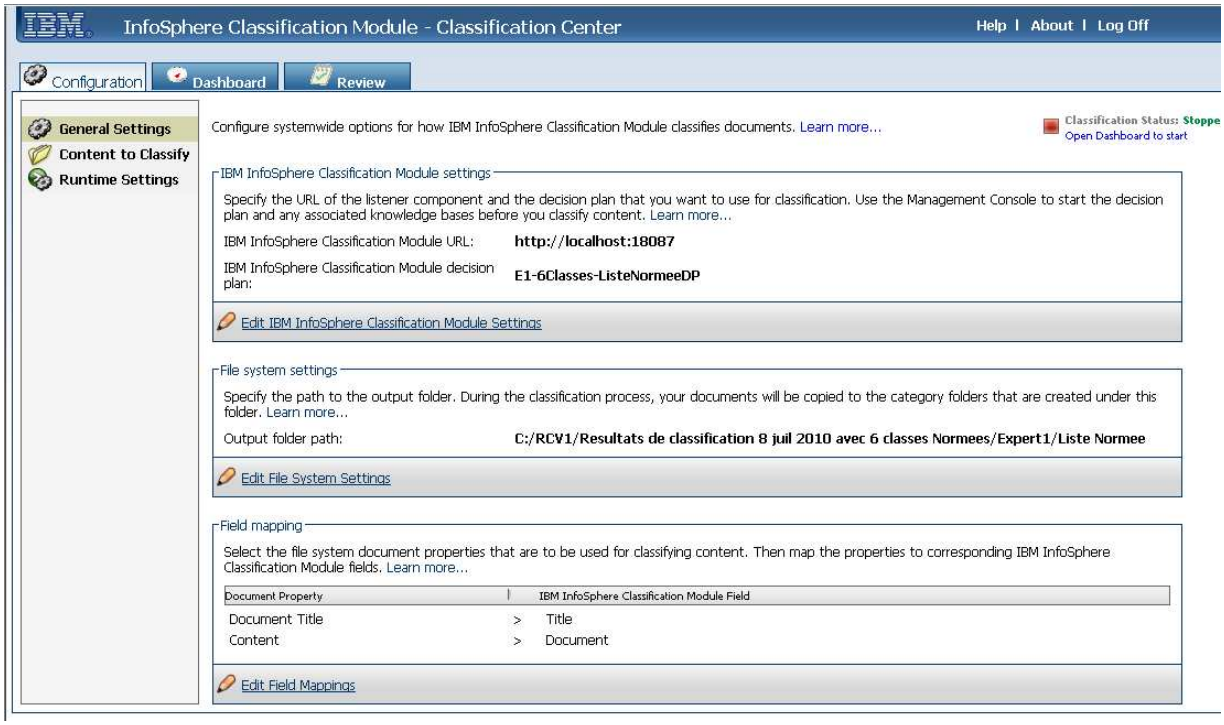


Figure 14: Centre de classification de document d'ICM

4.7.3 Récupération des résultats de la classification automatique

Une fois que la classification automatique des nouvelles est achevée, ICM dépose un résumé des résultats de la classification dans un fichier CSV (*Comma-separated values*) à partir duquel nous pouvons extraire, pour chaque nouvelle, les classes choisies ainsi que les scores correspondants (Figure 15).

A	B	C	D	E	F	G	H	I	J
<ICM_NVP key=""DP_Name""><![CDATA[E1-6CLASSES-LISTENORMEEDP]]></ICM_NVP>									
<ICM_NVP key=""DP_Version""><![CDATA[]]></ICM_NVP>									
<ICM_DP_Fired><![CDATA[New Group ^^ New Rule]]></ICM_DP_Fired>									
<ICM_NVP key=""Title""><![CDATA[XML00000001.xml]]></ICM_NVP>									
<ICM_DP_All_Changed_Names><![CDATA[FileSystem:Copy]]></ICM_DP_All_Changed_Names>									
<ICM_DP_Changed key=""FileSystem:Copy""><![CDATA[Notes]]></ICM_DP_Changed>									
<ICM_DP_Changed key=""FileSystem:Copy""><![CDATA[Structure]]></ICM_DP_Changed>									
<ICM_DP_Changed key=""FileSystem:Copy""><![CDATA[Acquisition]]></ICM_DP_Changed>									
<ICM_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.9961""><![CDATA[Notes]]></ICM_Match>									
<ICM_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.99373""><![CDATA[Structure]]></ICM_Match>									
<ICM_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.80657""><![CDATA[Acquisition]]></ICM_Match>									
<ICM_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.13892""><![CDATA[Price]]></ICM_Match>									
<ICM_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.0637""><![CDATA[Merger]]></ICM_Match>									
<ICM_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.02378""><![CDATA[Control]]></ICM_Match>									
<ICM_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.02357""><![CDATA[Reasons]]></ICM_Match>									
<ICM_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.02357""><![CDATA[Value]]></ICM_Match>									
<ICM_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.01959""><![CDATA[Financial]]></ICM_Match>									
<ICM_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.01959""><![CDATA[Liabilities]]></ICM_Match>									
<ICM_Final_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.9961""><![CDATA[Notes]]></ICM_Final_Match>									
<ICM_Final_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.99373""><![CDATA[Structure]]></ICM_Final_Match>									
<ICM_Final_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.80657""><![CDATA[Acquisition]]></ICM_Final_Match>									
<ICM_Final_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.13892""><![CDATA[Price]]></ICM_Final_Match>									
<ICM_Final_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.0637""><![CDATA[Merger]]></ICM_Final_Match>									
<ICM_Final_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.02378""><![CDATA[Control]]></ICM_Final_Match>									
<ICM_Final_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.02357""><![CDATA[Reasons]]></ICM_Final_Match>									
<ICM_Final_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.02357""><![CDATA[Value]]></ICM_Final_Match>									
<ICM_Final_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.01959""><![CDATA[Financial]]></ICM_Final_Match>									
<ICM_Final_Match kbname=""E1-6CLASSES-LISTENORMEEKB"" score=""0.01959""><![CDATA[Liabilities]]></ICM_Final_Match>									

Figure 15: Fichier events.csv contenant les résultats de la classification ICM des nouvelles

Le fichier CSV résultat de la classification automatique des nouvelles, contient la liste des étiquettes suivantes (basé sur un schéma XSD interne à ICM):

- ***ICM_NVP*** : est utilisé pour représenter la liste des données suivantes :
  - ***DP_Name*** : Nom du plan de décision dont le raisonnement a été utilisé pour classifier les nouvelles
  - ***DP_Version*** : Type de données véhiculées par le plan de décision
  - ***Title*** : Identifiant unique de la nouvelle traitée
- ***ICM_DP_Fired*** : nom de la (les) règle(s) déclenchée(s) pour le choix des classes
- ***ICM_DP_All_Changed_Names*** : Contient le type d'opération utilisée par ICM pour regrouper les nouvelles. Dans le cas de cet exemple « ***FileSystam:Copy »***, le code signifie que les nouvelles traitées seront copiées dans le(s) répertoire(s) correspondant(s) à(aux) la classe(s) choisie(s).
- ***ICM_Match***: Contient les classes utilisées dans la classification avec les scores correspondants à chacune des classes par rapport à la concordance (déduit grâce aux règles de décision) avec le contenu NLP de la nouvelle.
- ***ICM_DP_Changed*** : contient le nom des classes définitivement choisies grâce à leur score élevé pour classifier la nouvelle correspondante (ICM utilisera pour cela, le seuil spécifié par les règles de décision correspondantes)

4.8 Regroupement des résultats de la classification automatique selon les tableaux de contingence

L'analyse des résultats de la classification automatique des nouvelles par ICM prend en considération le fichier CSV contenant les résultats de la classification (voir la Figure 15 et les détails de la section ci haut).
Les données récupérées sont transformées de façon à mettre en évidence les éléments de la table de contingence de Sébastiani [29] (voir la Table 16).

Pour cela, nous déposons les résultats de la classification dans des tables de bases de données relationnelle MDB (*Microsoft Access database*) à l'intérieur desquelles des modules en VB calculent les TP, FP, FN et TN (Table 16) ensuite déduisent les mesures de performance nécessaires (Pour plus de détails sur les mesures utilisées, voir la section 3.4).

NewsID	Classes	Expert: Expert 2	Classificateur ICM		TN	FP	FN
13944	**Debt**	FAUX	FAUX	0	1	0	0
	Value	FAUX	FAUX	0	1	0	0
	Merger	VRAI	FAUX	0	0	0	1
	Acquisition	FAUX	VRAI	0	0	1	0
	Price	FAUX	FAUX	0	1	0	0
	Control	FAUX	FAUX	0	1	0	0
	Liabilities	FAUX	FAUX	0	1	0	0
	Reasons	FAUX	FAUX	0	1	0	0
	Structure	VRAI	VRAI	1	0	0	0
	Financial	FAUX	FAUX	0	1	0	0
	Notes	VRAI	VRAI	1	0	0	0
170233	**Debt**	FAUX	FAUX	0	1	0	0
	Value	VRAI	VRAI	1	0	0	0
	Merger	VRAI	VRAI	1	0	0	0
	Acquisition	FAUX	VRAI	0	0	1	0
	Price	FAUX	FAUX	0	1	0	0
	Control	FAUX	FAUX	0	1	0	0
	Liabilities	FAUX	FAUX	0	1	0	0
	Reasons	VRAI	VRAI	1	0	0	0
	Structure	VRAI	VRAI	1	0	0	0
	Financial	FAUX	FAUX	0	1	0	0
	Notes	VRAI	VRAI	1	0	0	0

Table 16: Exemple de la transformation du contenu d'un fichier CSV pour la classification

4.9 Comparaison et interprétation

La classification des échantillons de nouvelles par ICM sur la base respective de listes simple, hiérarchique et normée a permis d'obtenir des résultats différents nécessitant des mesures adaptées pour les comparer.

L'utilisation de différentes mesures (voir le chapitre 3) a permis d'obtenir la F-Mesure pour chaque résultat de classification comme cela est décrit dans les paragraphes qui suivent.

Dans ce qui suit, nous avons utilisé des tableaux de comparaison basés sur les différentes mesures décrites dans le troisième chapitre. La F-Mesure est notre mesure comparative car elle représente l'équilibre entre la rappel et la précision.

Les tables comparatives des sections 4.9.1, 4.9.2 et 4.9.3 ont été appliquées pour les 2 experts du domaines pour comparer la classification automatique d'ICM à celle manuelle des experts.
Les résultats varient selon qu'ICM a été entraîné sur la base du raisonnement de l'expert1 ou de l'expert2.

4.9.1 Utilisation des mesures classiques de Sébastiani

Pour le détail théorique de calcul de ces valeurs, se rapporter à la section 3.5.1.

Entraînement sur la base du raisonnement de l'expert$_i$ $i \in (1,2)$	**Liste simple**	**Liste hiérarchique**	**Liste normée**
Précision	PLS_i	PLH_i	PLN_i
Rappel	RLS_i	RLH_i	RLN_i
F-Mesure	**FLS_i**	**FLH_i**	**FLN_i**

Table 17 Calcul des mesures classiques de Sébastiani

Pour vérifier l'amélioration ou la diminution des performances de la classification automatique par rapport aux listes simple, hiérarchique et normée, on a comparé les valeurs **FLS_i** ,**FLH_i** et **FLN_i**.
La performance idéale est celle engendrant cette formule :
$FLN_i > FLH_i > FLS_i$

4.9.2 Utilisation des mesures élaborées et des mesures de Kiritchenko

Dans ce cas, nous avons utilisé les macro et micro moyennes ainsi que les mesures de Kiritchenko afin de présenter l'impact de la parenté entre les classes dans le cas d'une classification hiérarchique des nouvelles. Ainsi, pour chacune des 3 listes (simple, hiérarchique et normée) et chaque expert, nous avons calculé :

1. La micro-moyenne afin d'évaluer les performances globales de notre système de classification sans égard au poids des classes.
2. La macro-moyenne afin de faire un estimé moyen des performances globales de notre système de classification.
3. Les mesures de Kiritchenko afin d'analyser le comportement du classificateur lorsque la parenté des classes est prise en considération.

Étant donné que les valeurs obtenues sont complexes et puisque la F-Mesure est une base de raisonnement que nous avons adopté dans ce projet, nous avons résumé les mesures en ne gardant que les F-Mesures pour chaque liste et chaque expert.

Pour le détail théorique de calcul de ces valeurs, se rapporter à la section 3.5.2 et 3.6.

Entraînement sur la base du raisonnement de l'expert$_i$ i∈ (1,2)		Liste simple	Liste hiérarchique	Liste normée
Macro-Moyenne	**Précision**	MaPLS$_i$	MaPLH$_i$	MaPLN$_i$
	Rappel	MaRLS$_i$	MaRLH$_i$	MaRLN$_i$
	F-Mesure	**MaFLSi**	**MaFLH$_i$**	**MaFLN$_i$**
Micro-Moyenne	**Précision**	MiPLS$_i$	MiPLH$_i$	MiPLN$_i$
	Rappel	MiRLS$_i$	MiRLH$_i$	MiRLN$_i$
	F-Mesure	**MiFLSi**	**MiFLH$_i$**	**MiFLN$_i$**
Kiritchenko	**Précision**	KPLS$_i$	KPLH$_i$	KPLN$_i$
	Rappel	KRLS$_i$	KRLH$_i$	KRLN$_i$
	F-Mesure	**KFLSi**	**KFLH$_i$**	**KFLN$_i$**

Table 18 Calcul des mesures globales et de Kiritchenko

Pour vérifier l'amélioration ou la diminution des performances de la classification automatique par rapport aux listes simple, hiérarchique et normée, on a comparé les valeurs suivantes :

- Pour un estimé moyen des performances du système de classification, on a comparé les valeurs **MaFLS$_i$** , **MaFLH$_i$** et **MaFLN$_i$**.

  La performance idéale est celle engendrant cette formule :
  MaFLN$_i$ > MaFLH$_i$ > MaFLS$_i$

- Pour une évaluation des performances globales du système de classification, on a comparé les valeurs **MiFLS$_i$** , **MiFLH$_i$** et **MiFLN$_i$**.

  La performance idéale est celle engendrant cette formule :
  MiFLN$_i$ > MiFLH$_i$ > MiFLS$_i$

- Pour vérifier l'impact de la parenté et de la distance entre les classes, on a comparé les valeurs **KFLS$_i$** , **KFLH$_i$** et **KFLN$_i$**.

  La performance idéale est celle engendrant cette formule :
  KFLN$_i$ > KFLH$_i$ > KFLS$_i$

4.9.3 Utilisation de la mesure proposée

Afin de mieux comprendre la façon dont ICM a classifié les nouvelles sur la base d'une liste normée, la méthode proposée a été utilisée.

Le nombre de nouvelles bien, mal et mieux classifiées sont représentées dans le tableau type suivant (Pour le détail théorique de calcul de ces valeurs, se rapporter à la section 3.7).

Code	Signification	Totaux		
		Liste Simple Versus Liste Hiérarchique	**Liste simple Versus Liste normée**	**Liste hiérarchique Versus Liste normée**
AT	***Amélioration***	AT_{S-H}	AT_{S-N}	AT_{H-N}
AP	***Amélioration***	AP_{S-H}	AP_{S-N}	AP_{H-N}
NBCPE	***Amélioration***	$NBCPE_{S-H}$	$NBCPE_{S-N}$	$NBCPE_{H-N}$
NBC	***Bonne classification stable***	NBC_{S-H}	NBC_{S-N}	NBC_{H-N}
ACE	***Enrichissement***	ACE_{S-H}	ACE_{S-N}	ACE_{H-N}
ES	***Enrichissement***	ES_{S-H}	ES_{S-N}	ES_{H-N}
EA	***Enrichissement***	EA_{S-H}	EA_{S-N}	EA_{H-N}
ED	***Enrichissement***	ED_{S-H}	ED_{S-N}	ED_{H-N}
NMC	***Stabilité négative***	NMC_{S-H}	NMC_{S-N}	NMC_{H-N}
NPBC	***Stabilité négative***	$NPBC_{S-H}$	$NPBC_{S-N}$	$NPBC_{H-N}$
DP	***Diminution légère de la performance***	DP_{S-H}	DP_{S-N}	DP_{H-N}
DT	***Diminution grave de la performance***	DT_{S-H}	DT_{S-N}	DT_{H-N}

Table 19 Calcul de la mesure proposée

La mesure utilisée est celle décrite dans la section 3.7 et appliquée aux 3 cas de figure suivants :

- Liste simple versus liste hiérarchique :

$$Fraction_{S-H} = (AP_{S-H} + AT_{S-H}) / [(AP_{S-H} + AT_{S-H}) + (DP_{S-H} + DT_{S-H})]$$

La performance idéale est celle permettant à la fraction d'approcher le 100% :

$$Fraction_{S-H} \approx 1$$

- Liste simple versus liste normée :

$$Fraction_{S-N} = (AP_{S-N} + AT_{S-N}) / [(AP_{S-N} + AT_{S-N}) + (DP_{S-N} + DT_{S-N})]$$

La performance idéale est celle permettant à la fraction d'approcher le 100% :

$$Fraction_{S-N} \approx 1$$

- Liste hiérarchique versus liste normée :

$$Fraction_{H-N} = (AP_{H-N} + AT_{H-N}) / [(APH_{H-N} + AT_{H-N}) + (DP_{H-N} + DT_{H-N})]$$

La performance idéale est celle permettant à la fraction d'approcher le 100% :

$$Fraction_{H-N} \approx 1$$

L'équation finale idéale serait celle qui respecterait les 3 idéaux ci-haut.

4.10 Résultats finaux obtenus

La classification précédente des échantillons de nouvelles a été analysée et des mesures de performance ont été déduites (voir la section 4.9 pour plus de détails).

Des étapes intermédiaires ont été nécessaires (voir Annexe K pour plus de détails) pour aboutir aux résultats finaux suivants :

4.10.1 Résultats par rapport aux 2 experts

Voici les résultats finaux comparés à ceux des 2 experts du domaine :

4.10.1.1 Résultats d'ICM sur les mesures de base en comparaison avec les 2 experts

EXPERT 1	Liste Simple	Liste Hiérarchique	Liste Normée
Précision	0,5870	0,7156	0,8173
Rappel	0,8104	0,8414	0,8407
F-Mesure	**0,6809**	**0,7734**	**0,8288**

EXPERT 2	Liste Simple	Liste Hiérarchique	Liste Normée
Précision	0,6684	0,7351	0,7701
Rappel	0,7474	0,8261	0,8728
F-Mesure	**0,7057**	**0,7779**	**0,8182**

Table 20: Classes révisées - Mesures de Sébastiani suite à la classification des nouvelles

Nous avons constaté ce qui suit pour chaque expert :

1. La valeur enregistrée par la liste hiérarchique est plus importante que celle de la liste simple (exemple : 0,7734 par rapport à 0,6809 pour l'expert 1)
2. La valeur enregistrée par la liste normée est plus importante que celle de la liste hiérarchique (exemple : 0,8288 par rapport à 0,7734 pour l'expert 1)

On constate que l'amélioration se poursuit et devient plus importante lorsque les niveaux d'hiérarchie sont plus importants (1 seul niveau pour la liste simple, 2 niveaux pour la liste hiérarchique et 3 niveaux pour la liste normée). La parenté entre les classes ainsi que l'utilisation d'une ontologie spécifique au domaine ont permit d'améliorer les résultats d'une classification hiérarchique de nouvelles dans le cas des mesures classiques de Sébastiani.

4.10.1.2 Résultats selon les mesures élaborées et Kiritchenko

E X P E R T 1 :			
Mesures	***Liste Simple***	***Liste Hiérarchique***	***Liste Normée***
Macro-F-Mesure	***0,5165***	***0,5056***	***0,4950***
Micro-F-Mesure	***0,6809***	***0,7734***	***0,8288***
Kiritchenko-F-Mesure	***0,6809***	***0,8397***	***0,7828***

Table 21: Expert 1 - F-mesures pour 462 nouvelles avec une liste de classes révisées

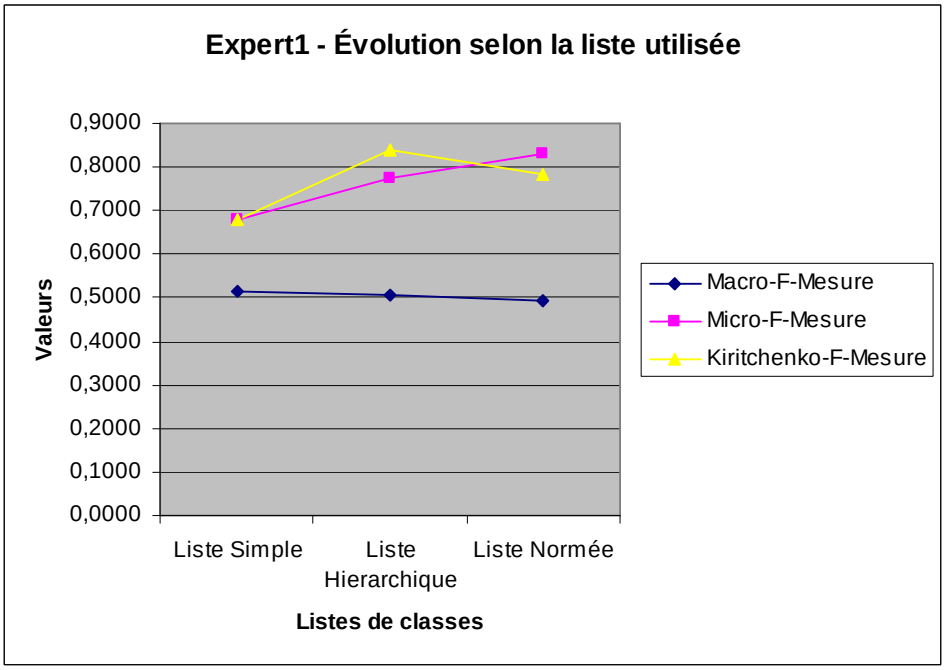


Figure 16: Évolution de la F-Mesure avec une liste normée à 6 classes pour l'expert 1

E X P E R T 2 :

Mesures	*Liste Simple*	*Liste Hiérarchique*	*Liste Normée*
Macro-F-Mesure	***0,4417***	***0,5159***	***0,5664***
Micro-F-Mesure	***0,7057***	***0,7779***	***0,8182***
Kiritchenko-F-Mesure	***0,7057***	***0,8593***	***0,8521***

Table 22: Expert 2 - F-mesures pour 462 nouvelles avec une liste de classes révisées

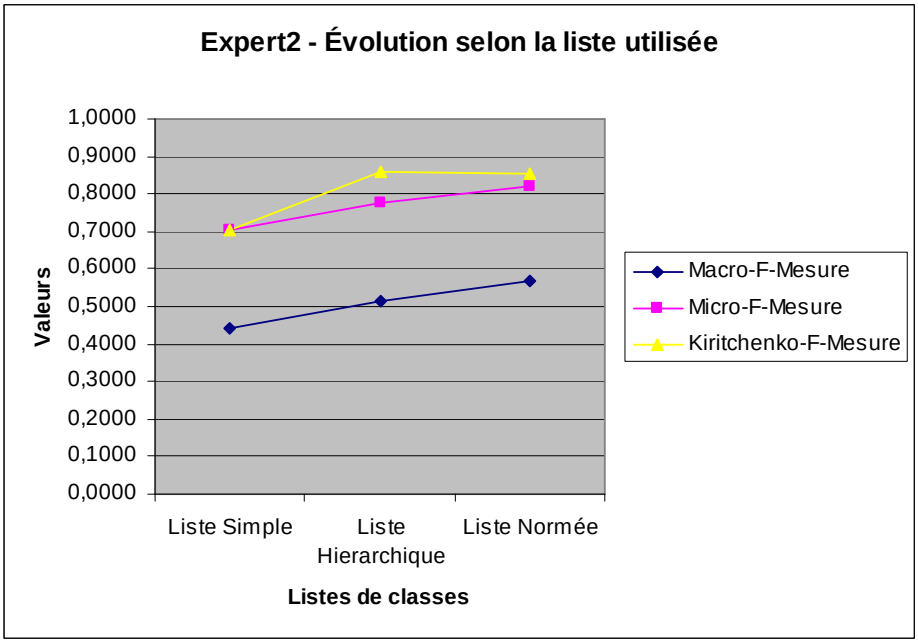


Figure 17: Évolution de la F-Mesure avec une liste normée à 6 classes pour l'expert 2

En analysant les graphes de la Figure 16 et de la Figure 17 nous notons une forme d'amélioration conséquente dans les mesures de micro et macro moyennes lorsque les niveaux d'hiérarchie de la liste de classes utilisée augmente (de la liste simple à la hiérarchique, et de la liste hiérarchique à la liste normée). Mais ce qui n'était pas prévu était la diminution des performances de la classification lorsque la liste normée est utilisée dans les mesures de Kiritchenko.

4.10.2 Résultats globaux

Lorsqu'on analyse la Table 22 on constate que l'expert2 a commis moins d'erreurs de classification puisqu'il enregistre de meilleurs scores de performance.
En se basant sur la logique de raisonnement utilisée par ICM et basée sur le raisonnement de l'expert 2, on constate les points suivants :

1. La parenté des classes a amélioré les résultats de la classification même si les mesures de Kiritchenko ont noté une baisse de performance dans le cas d'une liste normée.
2. Il existe un ensemble de nouvelles perturbateur (nouvelles mal classées) qui seront analysées dans la section suivante.

4.10.3 Analyse du raisonnement des experts

Afin de mieux comprendre la façon dont ICM a classifié les nouvelles sur la base d'une liste normée, nous avons utilisé la méthode enrichie des 4 cas de figure présentée à la section 3.7. En classifiant les nouvelles précédemment sélectionnées sur la base des listes simple, hiérarchique et normée, nous avons noté qu'ICM avait amélioré la classification de certaines nouvelles mais a mal classifié d'autres. La méthode des 4 cas de figure a permit de montrer de façon plus détaillée ces différents cas de classification en dressant des ensembles de nouvelles dont le lien commun concerne la façon dont ces dernières ont été classifiées. Ainsi, des ensembles de nouvelles bien, mal et mieux classifiées sont présentés dans la Table 23 pour l'expert 1 et la Table 24 pour l'expert 2.

Résumé Expert1	Liste Simple versus Liste Hiérarchique	Liste Simple Versus Liste Normée	Liste Hiérarchique Versus Liste Normée
Stabilité	219	209	224
Amélioration	235	214	204
Diminution	8	39	34
fraction	0,97	0,85	0,86

Table 23 Mesure proposée Expert 1

Résumé Expert2			
	Liste Simple versus Liste Hiérarchique	Liste Simple Versus Liste Normée	Liste Hiérarchique Versus Liste Normée
Stabilité	185	154	256
Amélioration	267	304	202
Diminution	10	4	4
fraction	0,96	0,99	0,98

Table 24 Mesure proposée Expert2

Si nous analysons les tables ci-haut, nous constatons qu'à la ligne fraction les résultats montrent le niveau de stabilité et d'amélioration ainsi que de diminution.
Si nous considérons que la stabilité ne peut influencer les performances d'un système de classification et que l'amélioration prouve son efficacité, nous avons calculé la fraction qui va permettre de comprendre le pourcentage des classifications améliorées par rapport au reste des classifications.
Cependant, les résultats montrent que le taux de diminution des performances modifie les résultats globaux en baissant le niveau de performance général. Ce constat s'applique surtout à l'expert 1 lorsque le passage vers une liste normée est effectué causant une forme biaisée des résultats.
En analysant mieux les classes choisies par l'expert 1, dans le cas des nouvelles en diminution de performance, une possibilité de mauvaise interprétation des nouvelles est apparue. Afin d'étudier cette piste de plus près, nous avons choisi de récolter toutes les nouvelles en diminution de performance, pour l'expert 1 et pour l'expert 2, et de les ré classifier manuellement par un expert professionnel du domaine.
Les résultats de cette classification manuelle experte seront, ensuite, injectés dans les classifications précédentes, en remplacement de celles biaisées, puis les mesures recalculées.

4.10.4 Résultats d'une re-classification corrective partielle – Mesure proposée

En re classifiant par l'expert 3 uniquement les nouvelles en diminution de performance, nous avons obtenu les résultats présentés dans les tables 25 et 26.

Les résultats enregistrés sont prometteurs et satisfaisants.

Résumé Expert1			
	Liste Simple versus Liste Hiérarchique	Liste Simple Versus Liste Normée	Liste Hiérarchique Versus Liste Normée
Stabilité	221	219	231
Amélioration	235	230	222
Diminution	6	13	9
fraction	0,98	0,95	0,96

Table 25 Re classification par l'expert3 des nouvelles en diminution de performance basées sur le raisonnement de l'expert1

Résumé Expert2	Liste Simple versus Liste Hiérarchique	Liste Simple Versus Liste Normée	Liste Hiérarchique Versus Liste Normée
Stabilité	186	159	257
Amélioration	267	299	202
Diminution	9	4	3
fraction	0,97	0,99	0,99

Table 26 Re classification par l'expert3 des nouvelles en diminution de performance basées sur le raisonnement de l'expert2

Si nous analysons les tables ci haut, nous constatons qu'à la ligne fraction les résultats montrent le niveau de stabilité et d'amélioration ainsi que de diminution. Cette dernière est beaucoup moins importante par rapport aux classifications précédentes (spécialement pour l'expert 1).
Cela prouve qu'ICM a pris de meilleures décisions en se basant sur un entraînement normé issu de l'accord des experts du domaine.
En fait, cela prouve que la diminution n'était pas causée par la liste normée utilisée mais plutôt par une mauvaise compréhension humaine. Cette marge d'erreur est une chose courante dans le domaine de la classification. Pourtant, le fait d'entraîner le classificateur commercial a permis de détecter cette anomalie et nous a permis de prouver de façon satisfaisante que l'amélioration d'une classification quelconque est surtout basée sur un bon entraînement et une liste de classes étudiée et organisée de façon ontologique.

La marge d'erreur est liée à la façon dont le raisonnement du classificateur a été orienté grâce à son entraînement.
Ce qui nous permet de prouver que les classifications sur la base d'une liste normée par rapport aux autres listes enregistre des performances timides mais néanmoins conséquentes.

CHAPITRE 5

Conclusion

5.1 Conclusion générale

À travers les résultats finaux obtenus, on a constaté ce qui suit lorsqu'il s'agit de la classification automatique des nouvelles sur la base d'une liste normée par rapport à d'autres listes simple et hiérarchique le utilisant le raisonnement d'experts du domaine :

1. Une liste normée a un certain effet sur la performance globale d'un classificateur
2. Une mesure plus détaillée est requise pour bien identifier les améliorations entre les listes
3. L'avantage majeur de la liste normée est l'augmentation de la stabilité du classement.

Le résultat de la stabilité est important du point de vue des utilisateurs (tels les analystes financiers dans les banques) qui eux interprètent les résultats. Pour ces utilisateurs, la stabilité qui augmente leur permet de conclure que leurs interprétations en finances seront cohérentes et pourront ainsi tirer des conclusions plus claires pour leurs investissements.
Ceci montre aussi que le but initial dans une classification automatique est la fiabilité des interprétations des résultats du logiciel.

5.2 Évaluation des contributions

5.2.1 Sommaire des résultats

Notre étude a débuté par le choix d'un échantillon de 1000 nouvelles classées manuellement par 2 experts en finances (voir la section 4.4).

Des mesures différentes ont été utilisées afin de varier l'angle de notre analyse en passant d'un angle global à un angle individualisé. Le fait de tester le système de classification sur un ensemble de données et de voir l'impact de cette classification sur les performances globales d'ICM concerne une analyse globale. Tandis que l'analyse en utilisant la mesure proposée va s'intéresser de près à des cas spéciaux de nouvelles en diminution de performance ce qui nous place à un angle individualisé pour juger les performances d'ICM selon tous les cas de figure de classification (Contradiction, Amélioration, Diminution et Stabilité de la performance).

Afin de comprendre les résultats d'une classification automatique, nous sommes passés des mesures classiques de Sébastiani vers des mesures hiérarchiques appliquées aux classificateurs multicritères et enfin vers les mesures hiérarchiques de Kiritchenko qui prennent en considération le lien de parenté entre les classes. Nous avons fait des constats variés qui nous ont obligés à adapter notre système de classification afin de corriger les anomalies constatées et d'obtenir des résultats de classification plus justes.

Initialement nous avons classifié un échantillon varié issu de la classification manuelle des nouvelles par les experts. Ensuite, nous avons classé automatiquement cet échantillon sur la

base d'une liste simple à 6 classes feuilles, puis d'une liste hiérarchique à 6 classes feuilles puis d'une liste normée à 14 classes feuilles. Notons que les classes feuilles choisies étaient dominantes dans le choix des 2 experts.

Aussi, nous avons considéré la F-Mesure comme étant une mesure de référence car elle permet d'équilibrer le poids entre les mesures du rappel et de la précision.

Les résultats de la classification étaient concluants lorsqu'il s'agissait de passer d'une classification sur la base d'une liste simple vers l'une ou l'autre des 2 autres classifications sur la base d'une liste hiérarchique et d'une liste normée. Mais le passage de la liste hiérarchique à la liste normée ne semblait pas améliorer les performances du classificateur (voir la section 4.10.1.1).

En analysant de plus près ces résultats grâce aux mesures de Kiritchenko, les résultats obtenus étaient positifs puisque la performance de la classification normée a enregistré une amélioration par rapport à l'hiérarchique. Cela signifie que les erreurs de classification sont plus faibles lorsque les classes choisies de façon erronée par ICM sont parentes avec les classes non choisies mais vraies (voir la section 4.10.1.2).
Mais, l'amélioration était faible et les mesures classiques semblent en contradiction avec celles de Kiritchenko même lorsque nous avons classé un autre échantillon plus important et plus riche que le premier (voir la section K.2 de l'Annexe K).

En analysant de plus près les listes de classes utilisées et non plus les nouvelles, on a noté un déséquilibre dans leur construction donnant lieu à des erreurs mathématiques et baissant de la sorte les performances du classificateur (voir la section K.2.3 de l'Annexe K). À partir de ce point, nous avons considéré par la suite que la classification des nouvelles obéissait à une probabilité différente lorsqu'on utilisait une liste simple et une liste hiérarchique par rapport à la liste normée. Car pour la liste hiérarchique, qui enregistrait des améliorations par rapport à la simple, une classe avait la probabilité 1/6 d'être choisie (il y a 6 classes feuilles) alors que la probabilité baisse (1/8) au niveau de la liste normée. Ce qui nous pousse à croire qu'une standardisation des listes des classes est nécessaire (voir la section K.3 de l'Annexe K).

Afin d'étudier cette nouvelle piste, la liste normée a été réduite aux mêmes classes utilisées dans les listes simple et hiérarchique en gardant les niveaux d'arborescence (voir la section k.3.1 de l'Annexe K). Une classification d'un nouvel échantillon plus représentatif de ces classes a été effectuée et les constats suivants ont été notés : Le fait de ne garder dans la liste normée que les classes feuilles similaires à celles des 2 autres classes, a permit d'améliorer de façon évidente les résultats de la classification des nouvelles sur la base d'une liste normée. En fait, on constate que l'amélioration se poursuit et devient plus importante lorsque les niveaux d'hiérarchie sont plus importants (1 seul niveau pour la liste simple, 2 niveaux pour la liste hiérarchique et 3 niveaux pour la liste normée).

Nous avons noté, une forme d'amélioration conséquente dans les mesures de micro et macro moyennes lorsque les niveaux d'hiérarchie de la liste de classes utilisée augmente (de la liste

simple à la hiérarchique, et de la liste hiérarchique à la liste normée). Mais ce qui n'était pas prévu était la diminution des performances de la classification lorsque la liste normée est utilisée dans les mesures de Kiritchenko.

Ce cas de figure nécessitant une analyse détaillée de chaque nouvelle ainsi que de chaque choix de classe, nous avons adopté une nouvelle méthode qui va se concentrer sur le raisonnement du classificateur plutôt que sur des calculs qui pourraient mettre de côté la valeur d'une classification améliorée et/ou enrichie pour une liste normée par rapport aux listes simple et hiérarchique. Cette méthode basée sur 4 cas de figure a permit de montrer de façon plus détaillée ces différents cas de classification en dressant des ensembles de nouvelles dont le lien commun concerne la façon dont ces dernières ont été classifiées.

Des ensembles de nouvelles bien, mal et mieux classées ont été dressés (voir la section 4.10.3 pour mieux comprendre la méthode). En prenant en considération la liste des nouvelles en diminution de performance, nous avons constaté que, même si elles ne sont pas nombreuses, ces dernières font en sorte de baisser le niveau de performance général du classificateur (voir la section 4.10.3). Ce constat s'applique surtout à l'expert 1 lorsque le passage vers une liste normée est effectué causant une forme biaisée des résultats (voir la Table 25).
En fait, cela prouve que la diminution n'était pas causée par la liste normée utilisée mais plutôt par une mauvaise compréhension humaine. Cette marge d'erreur est une chose courante dans le domaine de la classification. Pourtant, le fait d'entraîner le classificateur commercial a permis de détecter cette anomalie et nous a permis de prouver de façon satisfaisante que l'amélioration d'une classification quelconque est surtout basée sur un bon entraînement et une liste de classes étudiée et organisée de façon ontologique. La marge d'erreur est liée à la façon dont le raisonnement du classificateur a été orienté grâce à son entraînement.

Ce qui nous permet de prouver que les classifications sur la base d'une liste normée par rapport aux autres listes enregistre des performances timides mais néanmoins conséquentes si les conditions de classification utilisées étaient bien étudiées et les contradictions évitées.

5.2.2 Apport dans le domaine de l'optimisation des méthodes de la catégorisation automatique de textes

Comme apport au domaine de l'optimisation des méthodes de la catégorisation automatique de textes, une façon plus simple et plus centrée sur le raisonnement humain a été appliquée permettant ainsi de mieux comparer et de corriger les classifications automatiques avec celles des experts du domaine.

5.2.3 Comparaison avec des études similaires

Si nous comparons cette étude avec d'autres études similaires nous constatons que les domaines étudiés touchent, la plupart du temps, le médical, la recherche d'information en général [32] et le domaine du traitement langagier [2]. Ces derniers utilisent les ontologies à cause de la richesse des métadonnées s'y trouvant ainsi que le raisonnement logique utilisé qui rappelle les anciens systèmes experts.

Dans notre cas, on s'est concentrés sur le domaine des nouvelles financières qui attire de plus en plus de chercheurs en économie financière mais qui n'a pas eu encore toute l'attention nécessaire pour le faire avancer. Pourtant le volume de l'information financière augmente chaque année, ce qui impose une meilleure automatisation de la collecte et du traitement de cette dernière. Pour cela, une normalisation comptable de l'échange des informations est nécessaire.

Parmi les études similaires ayant utilisé les règles et les ontologies dans le domaines de la classification automatique, nous citons ce qui suit :

5.2.3.1 Étude comparative de He, J., et al.

« Catégorisation des connaissances en géni logiciels à travers l'utilisation d'une combinaison de SWEBOK et de catégorisation de textes » [16].

"*Categorizing Software Engineering Knowledge Using a Combination of SWEBOK and Text Categorization*"

Les auteurs de cette étude proposent d'utiliser une combinaison du premier spécimen de SWEBOK le SWEBOK1 (*Software Engineering Body of Knowledge*) et des techniques de catégorisation de textes afin de catégoriser des connaissances. L'épine dorsale du processus est le SWEBOK qui est utilisé comme taxonomie tandis que la catégorisation de textes fournit les algorithmes d'implémentation.

Les textes de simulation portent sur les actifs intellectuels et sont créés différemment selon leurs caractéristiques analysées. Ces connaissances sont améliorées ensuite avec de nouvelles fonctionnalités plus informatives générées grâce à l'utilisation d'ontologies créées manuellement. Ces dernières sont regroupées à l'intérieur d'un entrepôt de connaissances organisationnel.

Pour l'évaluation, un classificateur de textes SVM est utilisé pour catégoriser les objets de la connaissance. Il compare ses vecteurs documents d'entrée à tous les nœuds de la taxonomie SWEBOK puis renvoie le nombre souhaité de meilleures correspondances.

Les résultats expérimentaux ont prouvé que la méthodologie proposée permet d'automatiquement catégoriser les connaissances de géni logiciel explicitement et tacitement en même temps.

5.2.3.2 Étude comparative de Wang, T. et B.C.

« Classification de documents avec l'hiérarchie des sujets ACM » [15].

« Document Classification with ACM Subject Hierarchy »

Cette étude se base sur la construction d'un classificateur hiérarchique de textes pour la Bibliothèque numérique expérimentale CINDI[4]. Le système de classification construit utilise une procédure de catégorisation top-down en passant d'une catégorisation grossière à une raffinée. Le corpus utilisé dans l'expérimentation a été auto-généré et contient les articles de la science des ordinateurs archivés dans ACM DL.

[4] https://cindi.encs.concordia.ca

2 méthodes ont été adoptées permettant de re-classifier les documents à chaque niveau de l'arborescence.
Initialement un ensemble de classificateurs multi classes plats indépendants locaux au même nombre que les nœuds parents en utilisant un modèle top-down.
Un document est alors classifié en cascade en passant du classificateur racine de niveau 0 jusqu'aux feuilles (par raffinement).
Parmi les résultats observés, il a été noté que les classificateurs « *Naïve Bayes* » locaux surpassent les classificateurs de base Centroïdes.

5.2.3.3 Étude comparative de Hema Raghavan.

« L'apprentissage en tandem: un cadre d'apprentissage pour la catégorisation de documents » [1].

« Tandem learning: a learning framework for Document categorization »

l'apprentissage en tandem est un système d'apprentissage actif où le système reprend intelligemment les caractéristiques et les exemples pour l'enseignant afin de l'aider dans le processus d'étiquetage. Ainsi, la question sur la propriété d'une catégorie est spécifiquement limitée à la pertinence ou l'utilité d'une fonction permettant de déterminer si un objet appartient à une catégorie ou non. Dans cette étude, le terme «fonction de rétroaction" est utilisé afin d'impliquer l'assistance humaine dans la sélection d'une fonction. Le but est donc de prouver que les connaissances antérieures des utilisateurs (l'entraînement) sur les fonctions est utile pour la classification de textes.
Un algorithme pour l'apprentissage tandem débute avec un couple d'instances étiquetées, puis à chaque itération recommande à un utilisateur d'étiqueter les caractéristiques et les instances. Cet algorithme contient des méthodes pour intégrer la fonction de rétroaction en SVM (*Support Vector Machines*).
Un ensemble de mesures de difficulté a été conçu afin de permettre la capture de l'instance et de la complexité inhérentes au problème. La robustesse de ces mesures est prouvée à travers la corrélation entre l'instance et la complexité des fonctionnalités.

5.2.3.4 Étude comparative de Feldman, R

« Fouille de la littérature biomédicale en utilisant une analyse sémantique et des techniques de traitement automatique des langues naturelles » [2].

"*Mining the Biomedical Literature using Semantic Analysis and Natural Language Processing Techniques*"

Dans cette étude, l'intérêt est porté sur le « *text-mining* » appliqué à la littérature biomédicale. Ainsi, afin de faciliter la compréhension et la prédiction des processus biologiques complexes, des relations sont recherchées à travers des gènes, des protéines, des drogues et des maladies. Dans ce but précis, le système LitMiner[5] a été conçu en relation avec la découverte de connaissances et le Data Mining (KDD) Cup 2002 utilisé dans l'évaluation formelle du système.

[5] www.drugdiscoverytoday.com

Les modules utilisés dans l'extraction des entités et des relations sont implémentés en langage DIAL qui est spécifiquement désigné pour écrire des règles IE (*Information Extraction*).
Le system LitMiner a été développé en se basant sur un ensemble d'entraînement de 862 articles complets initialement taggués. Les résultats ont été probants surtout lorsque la F-Mesure a été calculée. Cela prouve la force des approches IE basées-règles (*rule-based*).

5.2.4 Pertinence et limites de cette étude

Étant donné que des études récentes ont démontré que les normes comptables avaient un impact sur le traitement des nouvelles financières, nous considérons que le fait d'utiliser une ontologie dans notre étude est justifiée et prouvée. La pertinence se résume surtout dans le besoin que le domaine des finances exprime.

Cependant, l'étude ne s'est concentrée que sur le domaine des fusions/acquisitions et a donc limité l'ontologie à une partie infime. Aussi, la classification manuelle étant ardue et nécessitant des moyens financiers et beaucoup de temps, l'échantillon manuellement classé a été restreint et, du coup, moins riche que ce qui était prévu au départ. Ce choix limité, mais nécessaire, nous a pourtant permis de traiter de façon individualisée chaque nouvelle et chaque classe et d'affiner au maximum les classes utilisées, les échantillons d'entraînement et de test et le raisonnement adopté lors de l'analyse des résultats de la classification automatique des nouvelles. Cette limite a aussi permit d'insérer dans cette étude un autre domaine de recherche à savoir, l'analyse du comportement des mesures hiérarchiques basées sur la classification hiérarchique multicritères prenant en compte des classes ayant des liens de parenté. C'est ainsi que nous avons prouvé que la parenté des classes réduisait le taux d'erreur lors de la classification hiérarchique et normée des textes.

5.3 Poursuite des travaux

Au plan théorique, notre étude a permit de déterminer la valeur relative des ontologies normées pour alimenter d'autres pistes de recherches prioritaires. Elle pourrait être utile aux chercheurs désireux de réduire la complexité de la base de connaissance utilisée.

5.4 Applications pratiques

5.4.1 Valeur ajoutée d'une classification normée en finances

Des études récentes semblent indiquer que les normes comptables auraient un impact sur la lecture et le traitement des nouvelles financières liées aux *fusion*s et *acquisitions*. Premièrement, il est important de comprendre que les nouvelles financières étant très succinctes, elles présentent une information comptable limitée, dont les montants pour une même variable (e.g., actif net, endettement, etc.) pouvant varier selon la norme utilisée [33]. De plus, puisque le coût d'une acquisition d'entreprise (i.e., le prix moyen évalué des actions multiplié par le nombre d'actions pour la prise de contrôle) est souvent affecté par les actifs intangibles (e.g., dépenses en R&D, marques de commerce, goodwill, etc.), il est important de bien relier les divers concepts comptables aux sujets pertinents qui les sous-tendent, ceux-ci pouvant affecter les valeurs rapportées et interprétées dans les nouvelles [34].

Enfin, les *fusion*s et *acquisitions* d'entreprises ont la particularité de donner une opportunité de refondre les états financiers de la nouvelle entité, et plusieurs options sont disponibles pour regrouper les variables comptables pour maximiser l'effet sur la valeur boursière de l'entreprise [35]. Donc dans l'ensemble, il est pertinent d'utiliser une ontologie normée pour structurer la hiérarchie ciblée pour une classification des nouvelles financières.

5.4.2 Systèmes d'analyse des nouvelles financières

Au plan des applications, nos résultats devraient servir à améliorer la performance liés au secteur des finances. Nous envisageons également des systèmes d'aide à la décision plus complexes, tels qu'un système de surveillance des marchés financiers permettant d'interpréter divers évènements affectant les sociétés cotées en bourse, dans le but de lier ces évènements à des prévisions des cours boursiers. Aussi, Nous pourrons améliorer l'analyse stratégique des transactions de *fusion* et *acquisitions* rapportées dans les nouvelles, surtout en identifiant la source précise dans les états financiers de l'information traitée, ainsi que les concepts auxquels ces données comptables sont associées.

5.4.3 Débouchées possibles pour d'autres domaines

Ceci étant, des problèmes plus profonds pourraient alors être mieux traités en se concentrant notamment sur les questions de la linguistique computationnelle ou encore des méthodes en recherche d'information (*Information Retrieval*).

ANNEXES

ANNEXE A

Engin de classification

A.1 IBM Classification Module : ICM

L'engin de classification utilisé dans cette recherche est un module du logiciel IBM OmniFind soit le module de classification ICM.

Intégrée à IBM FileNet®P8, la plate-forme d'entreprise ICM est utilisée par un certain nombre d'applications dans le but de classifier automatiquement des contenus. Parmi ces applications nous retrouvons, entre autres, l'automatisation des taxonomies et la classification de documents dans les systèmes ECM (*Enterprise Content Management*).

Incorporant la NLP dans le traitement des textes longs, et basé sur des règles, le système de classification ICM utilise aussi l'analyse des textes et l'apprentissage en temps-réel. Grâce aux librairies API clients (*Application Programming Interface*), ICM permet le développement rapide d'applications clients en usant de différents langages de programmation comme C, C++, Java™, C#, et Visual Basic, ainsi que des langages de script comme ASP en VBScript (*Active Server Pages*) [36].

A.2 Fonctionnalités du système ICM

ICM comporte une fonctionnalité très importante mettant en jeu le moteur de conception des relations (RME: *Relationship Modeling Engine*) afin de développer des applications permises RME [36]. Avec une telle fonctionnalité, les opérations suivantes sont possibles :

1. créer et configurer des bases de connaissance
2. faire correspondre du texte à une base de connaissance
3. fournir des feed-back à une base de connaissance
4. identifier le langage des textes
5. supporter des applications multilingues.

A.3 Étapes de création d'une base de connaissance

Les étapes de création d'une base de connaissance se déroulent de la façon suivante [37] (voir la Figure 18):

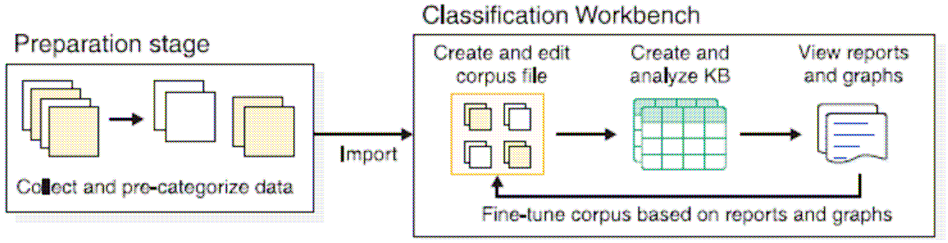


Figure 18: Liste de traitements sous le module de classification Workbench [38]

1. Importer des données devant servir à créer un fichier de corpus.

   Ces données peuvent avoir différents formats externes (CSV, XML, PDF, HTML, …) ou internes (ex : le corpus de Classification Workbench, le module de classification d'IBM, et les fichiers API 5.5 de l'engin de modélisation des relations RME)

2. Éditer et catégoriser les items de corpus
3. Créer et tester la base de connaissance en utilisant le corpus comme entrée
4. Évaluer la base de connaissance en utilisant des rapports et des diagnostiques graphiques
5. Améliorer les performances de la base de connaissance en ré-entraînant le corpus
6. La base de connaissance résultante est alors prête pour être utilisée avec les applications basées RME.

A.4 Fonctionnement du module de classification ICM

Le fonctionnement du module ICM peut être résumé dans le schéma ci-après :

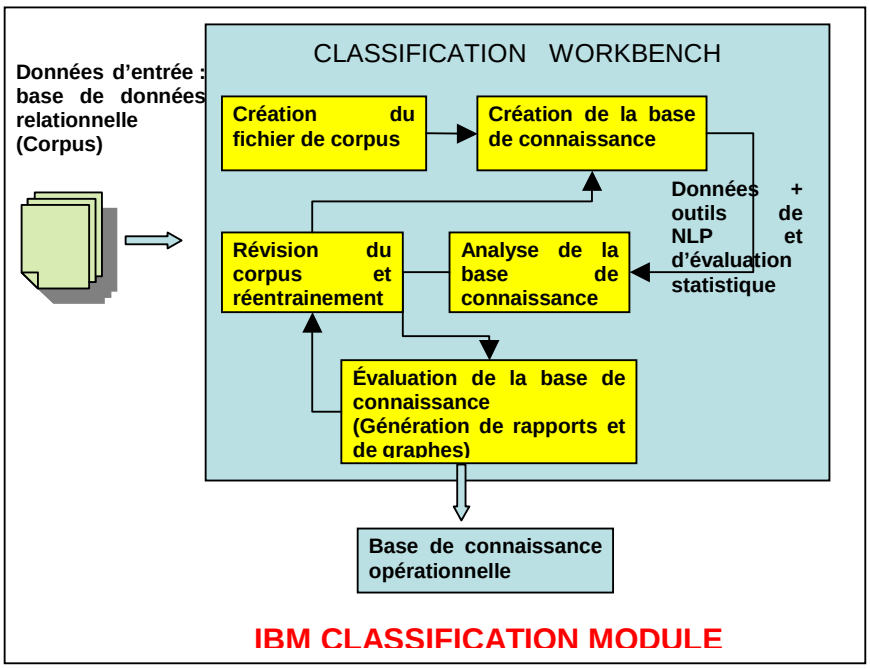


Figure 19: Fonctionnement d'IBM Classification Module

A.5 Fonctionnement schématisé de la classification de contenu dans le RME

Afin de créer un meilleur équilibre entre la variance et le biais[6], ICM s'appuie sur un algorithme unique et propriétaire le RME (voir la Figure 20)[37].

[6] L'une des questions les plus importantes en apprentissage machine est appelée le compromis variance-biais (*BIAS*). Ce compromis résulte du fait qu'un modèle qui augmente en complexité (son biais diminue) mais cela engendre aussi les variations énormes que subit l'identification paramétrique du modèle ce qui augmente la variance et vice versa.

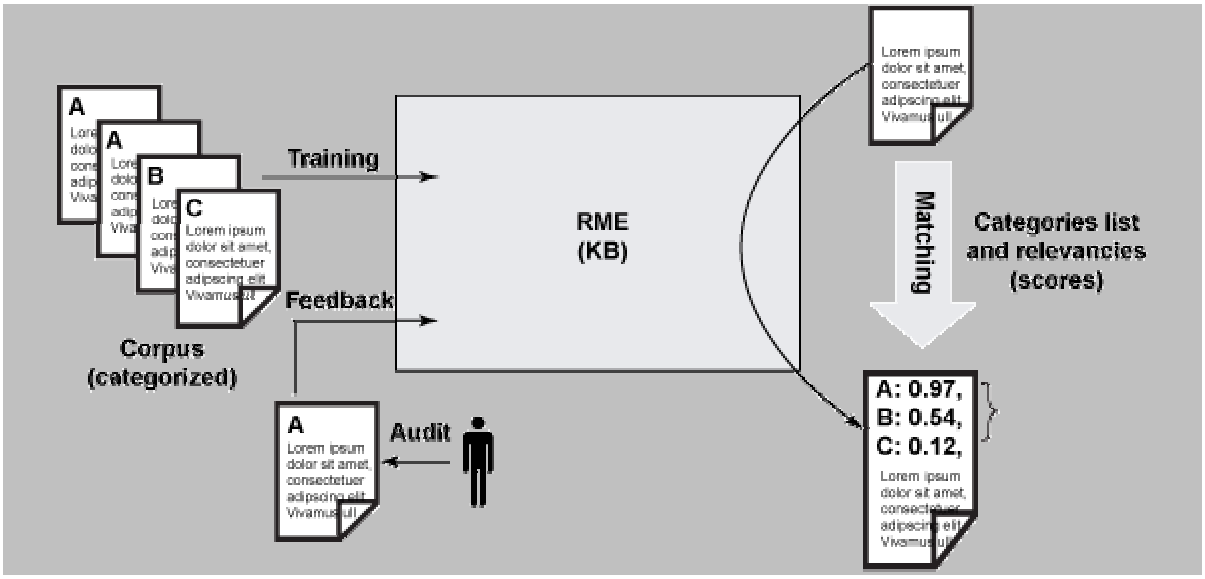


Figure 20: Fonctionnement de l'engin de modélisation des relations (RME) [37]

Partie droite du diagramme : l'engin de classification prend en entrée un document, l'analyse textuellement et propose en sortie un ensemble de catégories destinées à un système ECM.

Au centre du diagramme : l'engin de modélisation des relations d'ICM classifie le texte en entrée en respectant les 2 étapes suivantes :

1. Le moteur NLP extrait des concepts à partir de champs de texte libre (*free text*) pour générer un document en SML (*Semantic Modeling Language*)
2. Le moteur de modélisation sémantique génère les scores de pertinence (*relevancy scores*) en comparant le SML avec le contenu des catégories de la base de connaissance adaptative. Des catégories proposées sont retournées avec un score de pertinence représentant un facteur de confiance[7].

Partie gauche du diagramme : le contenu en phase de catégorisation taxonomiale est utilisé pour entraîner le système de façon continue. De cette façon, plus le corpus catégorisé grandit, plus les taxonomies sont affinées et d'autres catégories sont créées (ou modifiées).

[7] Le facteur de confiance (*confidence*) dérive du modèle de concepts et est comparé à un ensemble d'entrainement préalablement existant.

A.6 Les règles de décision

A.6.1 Définition des règles de décision

Les règles de décision permettent de déduire un ensemble d'actions en se basant sur certaines conditions initiales (voir la section A.6.2 pour plus de détails).
On peut les décrire selon leur utilité :

- Les règles de décision sont utilisées par le RME lors de l'analyse NLP des textes à classifier.
- Elles permettent de spécifier les seuils de scores pour rapprocher les textes aux définitions de classes utilisées.
- Elles permettent aussi de choisir des opérations apparentes aux choix de classes en déplaçant, par exemple, le texte à classifier dans un dossier représentant la classe choisie.

On peut utiliser une à plusieurs règles de décision par plan de décision et on peut activer ou désactiver une règle.

A.6.2 Création des règles de décision

Pour créer une règle de décision, les étapes suivantes sont nécessaires :

1. Définition des propriétés de la règle.
2. Définition d'un déclencheur pour la règle (*Trigger*).
3. Définition d'une ou plusieurs actions à exécuter lors du déclenchement de la règle.

Si on considère les concepts comme étant les classes et les documents à classer, et si on considère les déclencheurs comme le lien relationnel entre les concepts, on peut alors apparenter les règles de décision aux ontologies (Voir la section D.4 de l'annexe D traitant des ontologies).

A.6.3 Définition des propriétés de la règle

Les propriétés d'une règle servent à identifier ses comportements, entre autres, lorsqu'elle est déclenchée et lorsqu'une erreur survient lors de la classification. Le déclenchement des règles de décision ne se fait pas au hasard puisque ces dernières sont traitées de façon séquentielle l'une par rapport à l'autre.

Les 3 propriétés principales d'une règle de décision sont :

1. La propriété du comportement «*When triggered* » qui va contenir les actions à exécuter lorsque la règle est déclenchée.
2. Les actions que la règle exécutera lorsqu'une erreur est rencontrée « *On action error* ». Les actions peuvent concerner la continuité des autres actions ou l'arrêt total ou partiel.
3. L'état actif ou inactif de la règle. Cet état va permettre au déclencheur de la lancer dans le cas où les conditions de déclenchement sont satisfaites et que la règle est active.

A.6.4 Actions possibles d'une règle de décision

Les plans de décision d'ICM permettent d'exécuter automatiquement un ensemble d'actions (créées dans des règles de décision) lorsque les conditions d'un trigger sont rencontrées.

A.6.4.1 Actions prédéfinies

Ce sont des actions dont le but est de classifier du contenu dans des répertoires et utilisant des champs particuliers dont le nom est réservé dans un plan de décision. Le centre de classification les utilise lors de l'opération de classification de documents à travers le plan de décision spécifique.
Dépendamment du module utilisé, en plus des actions de déplacement ou de copie de document dans des répertoires désignés par les classes, on peut aussi avoir d'autres actions possibles permettant de faire ce qui suit, au déclenchement de ladite règle de décision :
DocumentClass : Action permise par l'outil FileNetP8 et consistant à créer une classe-document pour les documents à classifier.
FileNetP8_Metadata: Autre action permise par l'outil FileNetP8 consistant à ajouter un champ de métadonnée *FileNetP8_Metadata:metadata_field_name* aux documents. Le champ contiendra la valeur décrite par son nom.
FileSystem:Move : Action permise par l'outil *File System*. Permet de déplacer un document sous un répertoire particulier.

A.6.4.2 Actions avancées

Un ensemble d'autres actions personnalisées peuvent être créées à travers l'outil des actions avancées. On peut citer quelques-unes :
Add to content field: Ajoute une ou des valeur(s) à un champ de contenu.
Combine content fields: Combine un ensemble de champs de contenu pour former une variable temporaire dans le but de faciliter la recherche.
Extract pattern: Recherche un ou plusieurs modèles (pattern) dans un champ donné avant de les copier dans une variable temporaire.
Extract regular expression: Fait la même chose que l'action précédente mais recherche plutôt une expression régulière au lieu d'un pattern.
Match: Fournit la liste de documents qui correspondent (aux conditions du déclencheur) en utilisant la base de connaissance spécifiée.
Use simple hook: Utilise une application commune pour modifier le contenu d'un champ.

A.6.5 Les déclencheurs de règles - Triggers

A.6.5.1 Définition d'un trigger

Un déclencheur (trigger) permet de définir à travers une expression de contrainte les conditions nécessaires pour l'exécution des actions d'une règle quelconque (une règle au maximum).

On peut définir un déclencheur sur un ensemble d'actions prédéfinies suivantes :

Trigger always : Cette action va déclencher la règle sans aucune autre condition.

Trigger when fields contain specific words or phrases: le déclenchement de la règle se fera à la condition que certains mots ou phrases soient trouvés dans un champ de recherche spécifié

(généralement un champ avec contenu NLP) (voir l'étape 2 de la section 4.6.3.1).

Trigger when fields contain a substring : le déclenchement de la règle se fera à la condition qu'une chaine de caractères spécifique soit trouvée dans un champ de recherche spécifié.

Trigger when fields contain other strings : le déclenchement de la règle se fera à la condition qu'une chaine de caractères spécifique ou le contenu d'un champ est trouvé dans d'autres champs.

Trigger based on category scores : Permet de déclencher une règle en se basant sur la pertinence d'un élément par rapport à une classe. La règle devra utiliser, dans ce cas, les scores de pertinence.

Trigger based on the top-scoring category name : Le déclenchement de la règle sera basé sur la classe au plus grand score (donc la plus pertinente de la base de connaissance). Cette condition est utilisée dans l'étude actuelle.

Trigger based on word distance : Le déclenchement de la règle sera basé sur la proximité d'un ensemble de mots dans un champ de contenu.

A.6.5.2 Structure d'une expression de déclencheur

Les expressions de triggers utilisent une syntaxe particulière basée sur des valeurs booléennes de différentes façons :

```
(Boolean_condition) and not (Boolean_condition)
(Boolean_condition) or (Boolean_condition) or (Boolean_condition)
((Boolean_condition) and (Boolean_condition)) or not (Boolean_condition)
```

Les conditions booléennes disponibles sont : ***exists***, ***true***, ***numeric***, et ***string***. Les plus utilisées sont les suivantes :

1. Condition ***exists***

   La condition ***exists*** retourne la valeur *true* lorsqu'un champ de contenu existe dans le contenu d'un élément spécifié.

   Usage:

   ```
   $content_field_name exists
   exists $content_field_name
   exists $content_field_name[part_number]
   ```

   Part_number représente la valeur spécifique dont l'existence est vérifiée par la condition ***exists***.

2. Condition ***true***

   La condition ***true*** retourne la valeur *true* pour un pourcentage de temps spécifié.

   Usage:

   ```
   true/nn
   ```

   nn représente le pourcentage de temps permettant à la condition de retourner la valeur *true*.

3. Conditions ***numeric***

Les conditions ***numeric*** retournent la valeur *true* suite aux comparaisons numériques suivantes :

Equals
numeric_value = numeric_value

Does not equal
numeric_value <> numeric_value

Greater than
numeric_value > numeric_value

Greater than or equal to
numeric_value >= numeric_value

Less than
numeric_value < numeric_value

Less than or equal to
numeric_value <= numeric_value

Les valeurs numériques peuvent être simples ou extraites, entre autres, de la façon suivante :

$content_field_name
Extrait une valeur numérique à partir d'un champ de contenu de type: number.

var[temporary_numeric_variable_name]
Extrait une valeur numérique à partir d'une variable temporaire prédéfinie.

score('knowledge_base_name',category_name)
Extrait le score d'une catégorie à partir d'une base connaissance initialement spécifiée dans le plan de décision et contenant les champs *Match*.
Remarque: Un trigger qui retourne *true* chaque fois qu'un score de catégorie dépasse 70% est représenté par la syntaxe suivante:
score('knowledege_base_name' , 'category_name') > 0.70

score('knowledge_base_name',n)
Extrait le score d'une classe à partir d'une base de connaissances spécifiée initialement dans le plan de décision. n représente l'ordre de pertinence: exemple, n=3 va retourner le 3ème plus grand score.

4. Comparaisons de ***string***

Les conditions ***string*** retournent la valeur *true* lorsque 2 valeurs string équivalentes sont trouvées.

Usage :

string is string

Un string peut prendre une simple valeur ou une valeur extraite d'un champ de contenu.

Exemple :

$*content_field_name* (exemple, *$author* extrait John Doe)

5. Recherches de ***string***

   Les recherches de ***string*** retournent la valeur *true* lorsqu'un string est trouvé dans un champ de contenu.

   Usage:

   $*content_field_name contains string_constant*

6. Recherche de mots

   Une recherche de mot recherche les mots et/ou les phrases dans un champ de contenu. Le texte est déjà coupé en blocs d'unités par ICM (tokenized).

   Usage:

   $content_field_name : text_conditions

   lorsque la recherche se fait dans un seul champ de contenu.

   ou

   $__all__ : text_conditions

   lorsque la recherche se fait dans tous les champs.

   ou

   (text) and not (text)
   (text or text or text) and not (text)
   (text or (text and text)) and (text)

7. Fonction de classes

   Pour extraire une classe d'une base de connaissance en se basant sur son rang dans les résultats de l'entraînement (champ Match inséré dans la base de connaissance par l'entraînement) on utilise ***cat***.

   Usage :

   cat('knowledge_base_name',n)

   n représente l'ordre de pertinence.

8. Construction ***if-then-else***

   L'usage est :

   if (A) then (B) else (C).

   Permet de vérifier d'autres conditions pour déclencher une règle.

   Exemple:

   if (size($Categories) > 0) then ($Categories[1]) else ('<None>')

ANNEXE B

La classification hiérarchique de textes

B.1 Introduction

La classification de textes a pour objectif principal d'associer une ou plusieurs classes (catégories) à un document donné, en se basant sur son contenu.

Sachant que le contenu d'un document textuel peut varier du plus simple (courriel) au plus complexe (rapports), la plupart des classificateurs utilisent un traitement du langage naturel NLP pour le classifier mais sans considérer les relations existantes entre les classes. Ce genre de classification est dit *plat* car il ne suppose aucune relation structurelle qui pourrait lier les classes à utiliser [11].

Dans un classificateur plat, les classes sont traitées de façon égale et chaque document peut être classifié dans une classe ou dans un certain nombre de classes [15].

Note :
Les tâches de la classification de textes sont distinguées selon la cardinalité de l'ensemble des classes |C|. Ainsi, si |C|=2, la tâche de classification est dite binaire; Si |C|>2, la tâche de classification est dite multi-classes.

Aussi, si chaque document doit être assigné à exactement une classe, la tâche est dite à étiquette unique; Si chaque document peut être assigné à un nombre quelconque de classes allant de 0 à |C|, la tâche est dite multi-étiquettes.

Cardinalité de |C| / **Un document peut être assigné à**	**=2**	**>2**
Une classe au maximum	Classification binaire à étiquette unique	Classification multi-classes à étiquette unique
Un nombre quelconque de classes allant de 0 à |C|	Classification binaire à étiquettes multiples	Classification multi-classes à étiquettes multiples

Table 27: Table résumant les différentes tâches de classification

B.2 Les différentes approches dans le domaine de la classification plate de textes

Beaucoup d'approches en statistiques et en apprentissage machine se sont penchées sur la classification de textes parmi lesquelles on retrouve [17]:

1. Les modèles de régression multi-variés
2. Le plus proche voisin
3. Les modèles bayésiens probabilistes
4. Les arbres de décision
5. Les réseaux neuronaux
6. Les machines à vecteurs de support SVM « *support vector machines* »

Cependant, la plupart de ces approches utilisent un ensemble d'entraînement composé de documents pré-catégorisés (l'assignation document-classe est déjà faite et utilisée pour que le classificateur apprenne à classifier de nouveaux documents ne faisant pas partie de cet ensemble d'entraînement) et ignorent la structure hiérarchique de l'ensemble des classes [39]. Cela dit, la réalité montre que la plupart des systèmes d'information organisent leurs documents dans une hiérarchie de classes afin de faciliter la recherche par sujet. Cette façon d'organiser les documents a été ignorée, du moins jusqu'à la moitié des années 90, par la plupart des approches qui ont remplacé la hiérarchie des classes par un ensemble plat de classes.

B.3 La classification hiérarchique de textes

Le problème de la classification hiérarchique de textes a enfin été abordé en 1997 par Koller et Sahami [40]. Les classes étaient alors partiellement ordonnées de la classe la plus générique à la plus spécifique permettant de diviser un problème de recherche selon le sujet recherché à travers la dépendance entre les classes et l'ordre partiel les unissant. Grâce à cette approche, les résultats ont gagné en efficacité et en précision.

Les approches ont depuis beaucoup évolué se concentrant, globalement, autour de certains types de classifications hiérarchiques telles l'approche globale et l'approche locale.

B.3.1 Approche globale « big-bang »

De la même façon qu'une approche plate, un classificateur unique est construit traitant simultanément, mais de façon discriminatoire, toutes les classes d'une hiérarchie à la différence près que les relations entre les classes sont prises en considération [12].

Les approches considérées comme globales sont reliées, entre autres, à l'apprentissage par arbre de décision, les SVM, l'apprentissage probabiliste et les modifications hiérarchiques en apprentissage des règles d'association. Les recherches qui ont adopté cette approche sont nombreuses parmi lesquelles on peut citer : Weigend & al. 1999, Sasaki & Kita 1998 et Wang & al. 1999.

B.3.2 Approche locale « top-down level-based »

Dans cette approche, plusieurs classificateurs plats sont construits pour chaque nœud d'une hiérarchie [12]. Chacun de ces classificateurs travaille de haut en bas « *Top-down* » en choisissant initialement les classes de niveau supérieur les plus probables puis récursivement parmi les classes plus basses dans la hiérarchie descendantes de ces classes de niveau supérieur. Les documents sont classifiés dans des classes au niveau racine, ensuite dans chacune des sous-classes descendantes jusqu'à atteindre les classes feuille. Cette façon de faire permet de partitionner le problème en plusieurs sous-problèmes plus gérables.

Les approches dites locales font référence aux algorithmes d'apprentissage probabiliste, de réseaux neuronaux et SVM.

Les recherches ayant implémenté cette approche sont nombreuses mais les plus importantes sont celles d'Alessio & al. 2000, Dumais & Chen 2000, Koller & Sahami 1997 et Ruiz & Srinivasan 2002.

B.3.3 Autres approches

Il existe d'autres approches adoptées par la classification hiérarchique de textes qui se basent sur les méthodes probabilistes ou sur les techniques de *Clustering*, entre autres [13].

Dans les méthodes probabilistes, on retrouve une étude très intéressante qui concerne la technique *Shrinkage* de McCallum & al. 1998 qui permet de lisser les termes dans un document grâce à ses ancêtres [41]. Cette technique associe à chaque nœud de la hiérarchie une estimation de probabilité maximum basée sur les données associées à ce nœud. Ainsi, en estimant les distributions locales de chaque nœud et celles de ses parents hiérarchiques, les classes faiblement représentées peuvent être considérées d'une meilleure façon et les erreurs de classification sont significativement réduites [42].

Aussi, fait non négligeable, cette méthode est très utile lorsque les données d'entraînement sont rares spécialement dans le cas où de petits ensembles de documents sont assignés aux classes feuilles [11].

B.4 Définition de la structure de l'ensemble des classes dans une classification hiérarchique de textes

B.4.1 Le plan de classification

Lorsqu'on parle de structure hiérarchique adoptée par une classification hiérarchique de textes, on désigne le plan de classification correspondant et qui peut prendre l'une des deux formes suivantes [15]:

1ère forme : L'arborescence

L'arborescence contient les relations (parent-enfant) qui représentent la généralité ou la spécificité entre deux classes. Les feuilles de l'arbre représentent souvent les vraies classes et les nœuds internes sont surtout des classes virtuelles servant à continuer la classification :

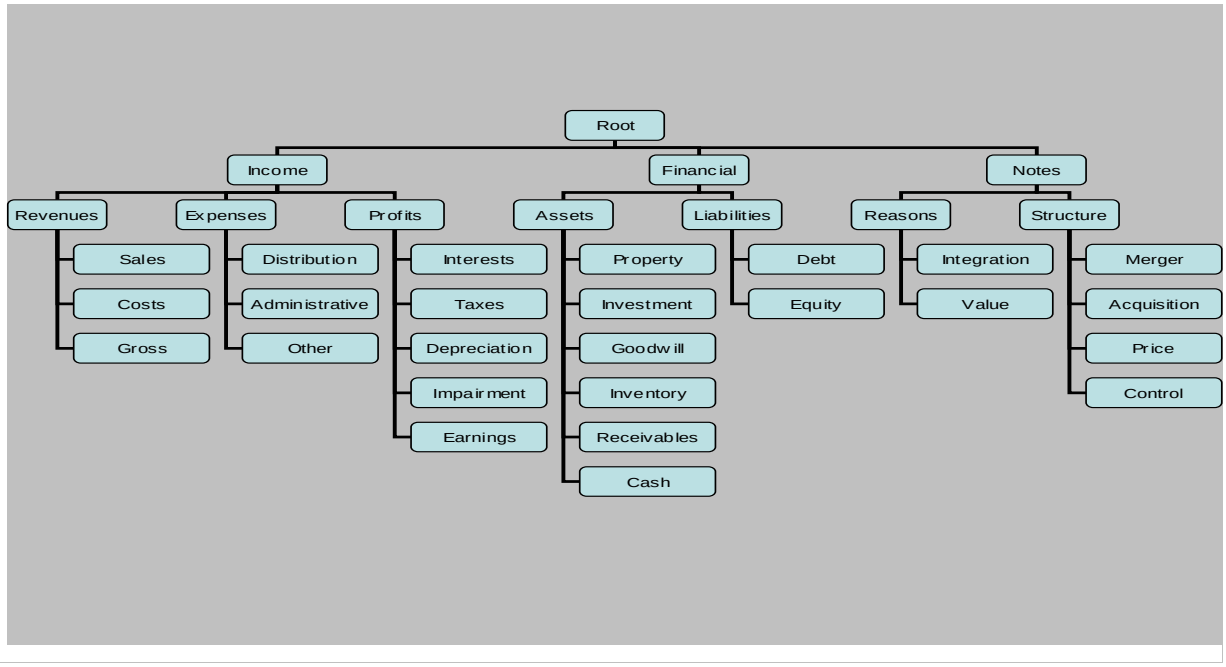


Figure 21: Ontologie du projet représentant l'arborescence de sujets financiers

2ème forme : Le graphe DAG « Directed Acyclic Graph »

Le graphe contient des arcs qui représentent la généralité ou la spécificité entre deux classes. Une classe peut avoir plus qu'un ancêtre directe (classe générale) ce qui implique que des documents peuvent être associés à une même classe mais à travers des chemins différents.

B.4.2 Les types d'information représentée par la hiérarchie des classes

Les classificateurs hiérarchiques de textes utilisent des méthodes différentes pour la classification des documents et l'évaluation des classifications par rapport aux classificateurs plats. Cette différence tient compte, entre autres, des types d'informations représentées par la hiérarchie des classes. Ainsi, la hiérarchie représente souvent des vocabulaires contrôlés[8] [43] sous la forme de taxonomies[9] ou de thésaurus[10]. Mais on peut aussi retrouver une hiérarchie sous la forme d'une ontologie qui, en plus d'organiser des entités, fournit la sémantique de ces dernières ainsi que celle des relations qui les lie.

[8] Selon la *Librairie Nationale du Canada*, un vocabulaire contrôlé établit une terminologie normalisée pour une utilisation dans l'indexation et la recherche d'information.

[9] Les taxonomies organisent des entités dans une hiérarchie.

[10] Un thesaurus fournit des relations entre des termes et en propose d'autres.

B.5 Formalisation de la classification hiérarchique de textes

B.5.1 Définition de la classification hiérarchique

La classification hiérarchique est une tâche qui assigne une valeur booléenne à chaque paire $\langle d_j, c_i \rangle \in D \times C$ avec [12] :

- D un domaine d'instances (ou de documents)
- $C = \{c_1, ..., c_{|C|}\}$ Un ensemble prédéfini de classes
- Les relations entre les classes de C sont décrites par une structure H correspondant à un ensemble partiellement ordonné tel que $H = \langle C, \leq \rangle$

B-5.2 Définition de l'ensemble fini partiellement ordonné

L'ensemble fini partiellement ordonné H est une structure telle que $H = \langle C, \leq \rangle$ avec :

- C un ensemble fini de classes
- $\leq \subseteq C \times C$ Est une relation sur C réflexive, antisymétrique et binaire transitive.

Ainsi pour chaque deux classes $p, q \in C$ avec $q<p$, $\nexists$ r : $r \in C$: q<r<p on dira alors:

- p est une classe parent de q
- q est une classe enfant de p
- Ancêtres(p)={ $q \in C$: $q \geq p$ } est l'ensemble des ancêtres de p
- Descendants(p)={ $q \in C$: $q \leq p$ } est l'ensemble des descendants de p
- Toutes les classes sans enfants sont des feuilles
- Toutes les classes ayant des parents et des enfants sont des classes internes

Aussi, pour chaque structure $H = \langle C, \leq \rangle$ on assume l'existence de la classe racine *Root(H)* qui est l'ancêtre de toutes les classes dans la hiérarchie.

B.6 Les mesures d'évaluation d'une classification hiérarchique de textes

B.6.1 Introduction

Les mesures appliquées aux classificateurs plats telles la précision et le rappel, ne conviennent pas à une classification hiérarchique dans la forme qu'on leur donne car elles ne prennent pas en considération les types d'erreurs liées à la mauvaise classification [12].

Ainsi, si une classe choisie est proche (parent ou frère) de la classe vraie, alors l'erreur est moins importante que si cette dernière est distante. Pour cette raison, la meilleure façon d'évaluer ce genre de classification est d'adopter une mesure hiérarchique qui respecte les trois conditions développées par Kiritchenko & al. 2006 [12].

Considérons l'exemple suivant :

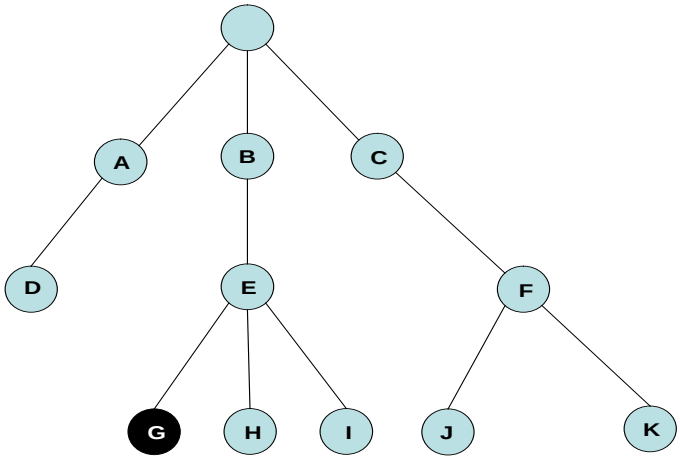


Figure 22: Exemple d'un arbre de classes appartenant à une classification hiérarchique. G est supposée être la classe vraie d'un document donné.

Les conditions proposées par Kiritchenko & al. sont les suivantes :

4. La mesure doit prendre en considération la classification partiellement correcte.

   Exemple
   En se basant sur la Figure 22, la classification d'un document sous la classe I devrait être moins pénalisante que celle sous la classe D en considérant G comme la classe vraie.

5. La mesure doit noter différemment les erreurs de distance selon le type de distance reliant la classe vraie de celle obtenue.

   Exemple
   En se basant sur la Figure 22, la classification d'un document sous la classe E devrait être moins pénalisante que celle sous la classe B en considérant G comme la classe vraie.
   « La mesure doit prendre en considération le nœud le plus proche selon son niveau »

6. La mesure doit pénaliser la mauvaise classification à un niveau supérieur.

   Exemple
   En se basant sur la Figure 22, la classification d'un document sous la classe I par rapport à la classe vraie G devrait être moins pénalisante que la classification sous C par rapport à la classe vraie A.

Ainsi, afin de respecter ces 3 conditions, les mesures doivent prendre en considération les ancêtres et les descendants d'une classe en plus de pouvoir mesurer la distance.

B.6.2 Formalisation

Soit une hiérarchie arborescente $H = \langle C, \leq \rangle$, on dénotera $HM(c_1 \mid c_2)$ l'évaluation hiérarchique de la classification d'un document $d \in D$ sous la classe $c_1 \in C$ sachant que $c_2 \in C$ est la vraie classe, alors, les trois conditions précédentes pourront être formalisées de la façon suivante :

Condition1 : Classification partiellement correcte : parenté des classes

Pour toute instance $(d, c_0) \in D \times C$,

Si Ancêtres(c_1)∩Ancêtres(c_0) ≠Φ **et** Ancêtres(c_2)∩Ancêtres(c_0) =Φ

Alors $HM(c_1 \mid c_0) > HM(c_2 \mid c_0)$

Condition2 : Erreurs de distance

Pour toute instance $(d, c_0) \in D \times C$,

Si c_1=Parent(c_2) **et** distance(c_1,c_0)>distance(c_2,c_0)

Alors $HM(c_1 \mid c_0) < HM(c_2 \mid c_0)$

Condition3 : Mauvaise classification à un niveau supérieur

Pour toutes instances $(d_1, c_1) \in D \times C$ et $(d_2, c_2) \in D \times C$,

Si
- distance$(c_1, c_1') =$ distance(c_2, c_2')
- Niveau(c_1)=niveau(c_2)+Δ
- Niveau(c_1')=niveau(c_2')+Δ

} Δ>0

- $c_1 \neq c_1'$, $c_2 \neq c_2'$
- avec : niveau(x) est la longueur du chemin de la racine au nœud x

Alors $HM(c_1' \mid c_1) > HM(c_2' \mid c_2)$

Comme les mesures de la classification plate ne respectent aucune de ces conditions et que les autres mesures (telles les mesures basées distance) ne respectent qu'une partie des conditions, une nouvelle mesure hiérarchique devait être élaborée.

Afin de répondre aux conditions précitées, une mesure hiérarchique particulière hF exploitant le rappel et la précision non seulement au niveau de chaque exemple mais surtout au niveau des ancêtres de chacun d'eux a été présentée par Kiritchenko en 2005 [43] dans sa thèse en fournissant la preuve que cette mesure satisfait les trois conditions pré-requises pour l'évaluation d'un classificateur hiérarchique à la structure arborescente ou graphique. La mesure hF utilise les notions d'ancêtres pour calculer les erreurs de classification [44]. Formellement, en considérant une classification hiérarchique multi-étiquettes, on peut définir la mesure d'évaluation hF de la façon suivante [12]:

Pour toute instance (d_i,C_i) classifiée sous le sous-ensemble $C_i^{'}$ avec $C_i^{'} \subseteq C$, $d_i \in D$, $C_i \subseteq C$, on aura Les micro-moyennes hP et hR telles que :

$$hP = \frac{\sum_i \left| \text{Ancêtre}(C_i) \cap \textit{Ancêtre}(C_i^{'}) \right|}{\sum_i \left| \text{Ancêtre}(C_i^{'}) \right|}$$

$$hR = \frac{\sum_i \left| \text{Ancêtre}(C_i) \cap \textit{Ancêtre}(C_i^{'}) \right|}{\sum_i \left| \text{Ancêtre}(C_i) \right|}$$

La combinaison des deux valeurs hP et hR permet de calculer la F-Score (hF) :

$$hF_\beta = \frac{(\beta^2 + 1) hP.hR}{\beta^2 hP + hR}, \ \beta \in [0,+\infty]$$

Afin de donner le même poids à la précision et au rappel, on utilise souvent $\beta = 1$.

ANNEXE C

Mesures d'évaluation basées sur une classification non hiérarchique

C.1 Mesures d'évaluation standards pour la classification binaire

Afin de mesurer les performances de la classification, le classificateur est considéré être capable de prendre les bonnes décisions de classification [45]. Toutefois, comme le problème de classification est non formalisable [29], le classificateur est alors évalué de façon expérimentale et non analytique.

Dans les tâches de classification binaire nous utilisons souvent les mesures d'évaluation standards de *précision* et de *rappel (recall*) inspirées des notions de la recherche d'information (*information retrieval*) classiques. Basées sur un jugement binaire : pertinent ou non pertinent, le rappel et la précision s'expriment de la façon suivante :

► La *précision* (degré de certitude) représente la probabilité conditionnelle que si un document d_j, pris au hasard, est classifié sous une catégorie c_i alors cette décision est correcte [29] :

$$Précision = Probabilité\ (d_j\ classé\ sous\ c_i = décision\ correcte) \quad (7)$$

► Le *rappel* (degré de complétude) représente la probabilité conditionnelle que si un document d_j, pris au hasard, devrait être classifié sous la catégorie c_i, alors cette décision est prise [29]:

$$Rappel = Probabilité\ (d_j\ doit\ être\ classé\ sous\ c_i = décision\ prise) \quad (8)$$

C.2 La table de contingence de Sébastiani

Selon Sébastiani [29], nous pouvons déduire une table de contingence (d'un sous-ensemble test donné) pour chaque document d_j classifié sous la catégorie c_i.

Ainsi, dans la Table 28 ci bas, nous considérons la pertinence du classement d'un document d_j dans la catégorie c_i selon les valeurs suivantes :

- les positifs vrais et les négatifs vrais représentent la bonne classification
- les positifs faux représentent les erreurs de commission (document classifié dans la mauvaise catégorie)
- les négatifs faux représentent les erreurs d'omission (document non classifié dans la bonne catégorie).

Catégorie c_i		Jugements des experts (prédiction)	
		OUI (Pertinent)	**NON** (Non pertinent)
Jugements du classificateur (état actuel)	**OUI**	$\mathbf{TP_i}$ Positif vrai	$\mathbf{FP_i}$ Positif faux
	NON	$\mathbf{FN_i}$ Négatif faux	$\mathbf{TN_i}$ Négatif vrai

Table 28: Table de contingence pour le jugement des résultats de classement

À partir de ces valeurs, nous calculerons les mesures de précision et de rappel de la façon suivante :

$$\text{Précision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Rappel} = \frac{TP}{TP + FN} \tag{10}$$

où $TP = \sum_i TP_i$, $FP = \sum_i FP_i$ et $FN = \sum_i FN_i$.

De façon textuelle, nous pouvons décrire ces 2 mesures de la façon suivante :

$$\textit{Précision} = \frac{\textit{Nombre de documents pertinents retrouvés}}{\textit{Nombre de documents pertinents}} \tag{11}$$

$$\textit{Rappel} = \frac{\textit{Nombre de documents pertinents retrouvés}}{\textit{Nombre de documents retrouvés}} \tag{12}$$

Une valeur élevée pour chacune des deux mesures est souhaitable mais, ces deux dernières sont liées de façon conflictuelle car dès que le système tente d'augmenter le « rappel » en augmentant le nombre de documents pertinents retrouvés, la précision se réduit car quelques documents non pertinents peuvent s'y retrouver (voir la Figure 23).

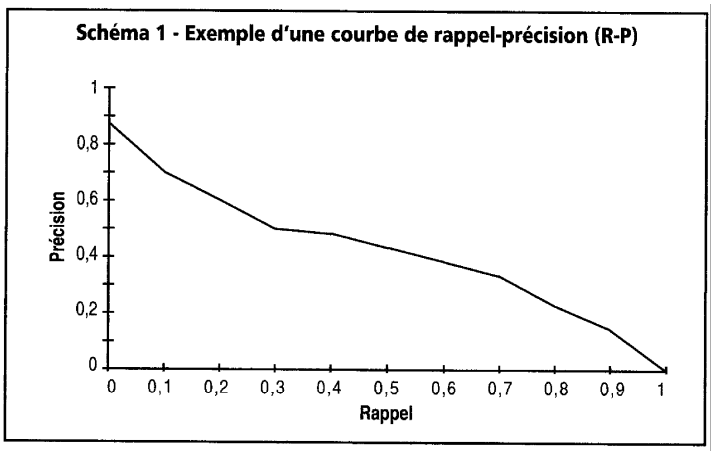


Figure 23: Exemple du conflit existant entre les mesures de précision et de rappel [46].

D'autres mesures sont utilisées dans l'évaluation des classificateurs binaires. Les plus connues sont *fallout*, *accuracy* et *error* qui s'expriment de la façon suivante :

$$Fallout = \frac{FP}{TN + FP} \tag{13}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{14}$$

$$Error = \frac{FP + FN}{TP + FP + FN + TN} \tag{15}$$

Cependant, les mesures *fallout* et *accuracy* ne sont pas sensibles aux catégories rares, et la mesure *error* (ainsi que *accuracy*) n'est pas adéquate lorsque le nombre de catégories est grand et que la moyenne des catégories par document est petite [47].

Une autre mesure de performance reconnue est la *F*1_measure. Définie par van Rijsbergen [30] elle sert à balancer les valeurs du rappel et de la précision. Sa forme générale introduit un paramètre β qui met en exergue la pondération différente du rappel (r) et de la précision (p) :

$$F\_Measure_{\beta}(r, p) = \frac{(\beta^2 + 1) pr}{\beta^2 p + r} \tag{16}$$

Nous désignerons par la suite cette mesure de façon plus simple par F_{β}. Le paramètre β est considéré comme un degré relatif d'importance attribué au *rappel* et à la *précision*. Lorsque β équivaut à la valeur 0 (zéro), F_{β} correspond alors à la *précision*. Lorsque β tend à l'infini, F_{β} correspond au *rappel*. Afin d'attribuer une importance égale au *rappel* et à la *précision*, β est souvent remplacée par la valeur 1 [29] et F_1 prend alors la forme suivante :

$$F_1(r, p) = \frac{2pr}{p + r} \tag{17}$$

C.3 Exemple de calcul de mesures de performance grâce aux tables de contingence

Dans l'exemple suivant, nous allons calculer les mesures de performance simples grâce à l'utilisation des tables de contingence décrites par « *Fabrizio Sébastiani* ».

Nous disposons d'un corpus initial Ω pour entraîner un classificateur Φ quelconque (pour plus de détails sur la formalisation de la classification en général, se référer à l'annexe C) :
$\Omega = \{d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}\} \subset D$ Avec un pré-classement sous les catégories $C = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7\}$ représenté par la table suivant :

Documents de Ω / Catégories	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
c1		x		x	x					
c2			x			x				
c3	x									
c4							x			
c5									x	x
c6								x		
c7										

Table 29: Documents pré-classifiés

Supposons que l'ensemble d'entraînement choisi est $\Omega_1 = \{d_1,d_2,d_3,d_4,d_5\}$ et que l'ensemble test est $\Omega_2 = \{d_6,d_7,d_8,d_9,d_{10}\}$. Le classificateur Φ sera donc entraîné en observant la façon avec laquelle des documents de Ω_1 ont été classés et sera testé en lui soumettant chaque document de Ω_2. Maintenant, supposons que le classificateur Φ a pris les décisions de classification suivantes pour les documents de Ω_2:

Documents de Ω_2 / Catégories	d_6	d_7	d_8	d_9	d_{10}
c2	x	x			
c5				x	
c6			x		
c7					x

Table 30: Décisions du classificateur

En comparant les deux tables ci haut, nous pouvons remplir la table de contingence pour chacun des documents tests de la façon suivante :

Document d6			
Catégorie c2		**Pertinence**	
		OUI	**NON**
Décision du classificateur	**OUI**	**1**	**0**
	NON	**0**	**0**

Document d6			
Catégories c1,c3,c4,c5,c6,c7		**Pertinence**	
		OUI	**NON**
Décision du classificateur	**OUI**	**0**	**0**
	NON	**0**	**1**

Document d7			
Catégorie c2		**Pertinence**	
		OUI	**NON**
Décision du classificateur	**OUI**	**0**	**1**
	NON	**0**	**0**

Document d7			
Catégorie c4		**Pertinence**	
		OUI	**NON**
Décision du classificateur	**OUI**	**0**	**0**
	NON	**1**	**0**

Document d7			
Catégories c1,c3,c5,c6,c7		Pertinence	
		OUI	NON
Décision du classificateur	OUI	0	0
	NON	0	1

Document d8			
Catégorie c6		Pertinence	
		OUI	NON
Décision du classificateur	OUI	1	0
	NON	0	0

Document d8			
Catégories c1,c2,c3,c4,c5,c7		Pertinence	
		OUI	NON
Décision du classificateur	OUI	0	0
	NON	0	1

Document d9			
Catégorie c5		Pertinence	
		OUI	NON
Décision du classificateur	OUI	1	0
	NON	0	0

Document d9			
Catégories c1,c2,c3,c4,c6,c7		Pertinence	
		OUI	NON
Décision du classificateur	OUI	0	0
	NON	0	1

Document d10			
Catégorie c7		Pertinence	
		OUI	NON
Décision du classificateur	OUI	0	1
	NON	0	0

Document d10			
Catégorie c5		Pertinence	
		OUI	NON
Décision du classificateur	OUI	0	0
	NON	1	0

Document d10			
Catégorie c1,c2,c3,c4,c6		Pertinence	
		OUI	NON
Décision du classificateur	OUI	0	0
	NON	0	1

En calculant la somme de chaque TP_i, FP_i, FN_i et TN_i de tous les documents tests selon les catégories du corpus, nous obtenons ce qui suit :

TP = 3 FP = 2 FN = 2 TN = 28

ANNEXE D

Les ontologies et les taxonomies

D-1 Historique

Bien que son origine remonte aux travaux en métaphysique d'Aristote, l'ontologie a été reprise par l'Intelligence Artificielle pour faire évoluer les recherches en représentation des connaissances [48]. Ainsi, la piste de la représentation explicite des connaissances a ouvert la voie à la représentation déclarative de ces dernières. Les systèmes à base de connaissance prirent alors l'envol durant les années 80 confortés par les avancées dans le domaine des systèmes experts. Ces bases de connaissance étant complexes ne permettaient pas leur réutilisation, du moins de façon automatique. Cela a donné l'idée à des chercheurs de collaborer au sein du projet « *Knowledge Sharing Effort* » initié par la DARPA « *Defense Advanced Research Projects Agency* »[11] pour donner forme à la « Représentation explicite du sens » qui sera appelée par la suite « Ontologie ».

D-2 Définitions générales

« En Intelligence Artificielle, une ontologie réfère à un artefact d'ingénierie ' Engineering Artifact' constitué d'un vocabulaire spécifique utilisé pour décrire une certaine réalité, plus un ensemble d'hypothèses explicites concernant le sens voulu des mots du vocabulaire » [49].

Dans une ontologie formelle, cet ensemble d'hypothèses implique que les mots du vocabulaire apparaissent comme des noms de prédicats (appelés concepts et relations) binaires ou unaires.

Selon Gruber (1993) [50], si on considère qu'un travail de conceptualisation est nécessaire avant de bâtir une ontologie, on peut alors définir cette dernière de la façon suivante :

« Une ontologie est une spécification explicite d'une conceptualisation ».

Cependant, et étant donné que les ontologies sont exprimées sous forme logique, et que la conceptualisation est souvent partiellement formalisée dans un cadre logique [32], alors on peut définir les ontologies de façon plus affinée. Ainsi Guarino (1995) considère les ontologies de la façon suivante :

« Les ontologies sont des spécifications partielles et formelles d'une conceptualisation ».

D-3 Définition opérationnelle

Une définition basée sur l'approche opérationnelle a été fournie par Wikipédia et décrit les ontologies de la façon suivante :

[11] Selon Wikipédia du 19 septembre 2009, la DARPA « agence pour les projets de recherche avancée de défense » est une agence du Département de la Défense des Etats-Unis qui se spécialise dans les nouvelles technologies à usage militaire. La DARPA est à l'origine, notamment, du lancement du réseau informatique ARPANET (*Advanced Research Projects Agency Network*) qui est le précurseur d'Internet.

« Une ontologie est comme un 'réseau sémantique' qui regroupe un ensemble de concepts décrivant complètement un domaine. Ces concepts sont liés les uns aux autres par des relations taxonomiques[12] d'une part, et sémantiques d'autre part »[13].

Selon les ontologistes, les concepts et les relations sont les primitives cognitives de base d'une ontologie [32]. Ainsi, une ontologie peut traduire les connaissances grâce aux éléments suivants : Les concepts, les relations, les axiomes et les instances.
Dans le cas le plus simple, une ontologie décrit une hiérarchie de concepts reliés par des relations de subsomption; Dans des cas plus sophistiqués, des axiomes adaptés sont ajoutés afin d'exprimer d'autres relations entre les concepts et de contraindre leur interprétation attendue.

D.3.1 Les concepts

Un concept (ou terme) est une entité qui peut représenter, entre autres, un objet (ex : matériel) ou une idée. Il est composé de trois parties (voir la Figure 24) :

1. Un (des) terme(s)
2. Une notion (concept sémantique) : également connue sous le nom d'« intension » du concept, elle représente la sémantique du concept exprimée avec un ensemble de propriétés (ex : généricité, rigidité,…), attributs, règles et contraintes.
3. Un ensemble d'objets (concept formel) : Également connu sous le nom d'« extension » ou « réalisation » du concept. Il regroupe les instances du concept (objets manipulés) (voir la Figure 24 et la Figure 25).

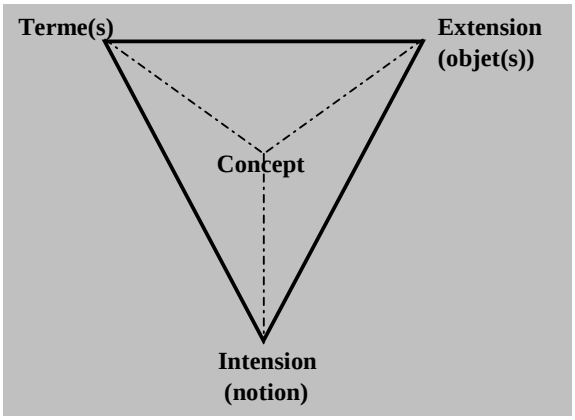


Figure 24: Triangle sémantique d'Ogden et Richards (1923)

[12] Hiérarchisation des concepts

[13] Wikipédia 15 septembre 2009. Lien : http://fr.wikipedia.org/wiki/Ontologie_(informatique).

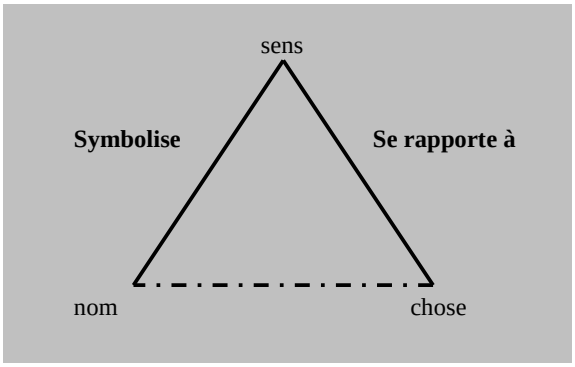
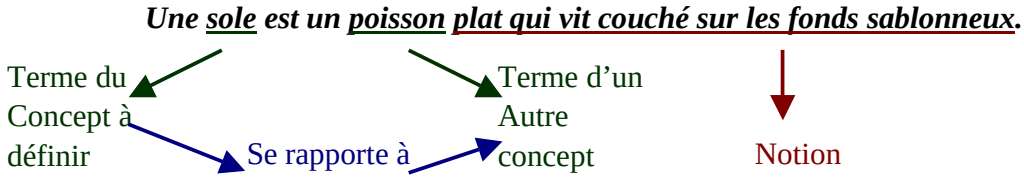


Figure 25: Triangle sémantique présenté schématiquement par Tamba 1991

Les concepts peuvent être classifiés selon les dimensions suivantes (Gomez-Perez A, 1999) :

- Niveau d'abstraction
- Atomicité
- Niveau de réalité

Soit un exemple pris du Larousse :



D.3.2 Les relations

Unissent les concepts et sont caractérisées par un (des) terme(s) et une signature. Elles incluent les associations suivantes :

- Sous-classe-de
- Partie-de
- Associée-à
- Instance-de

Les fonctions sont un cas particulier des relations.

D.3.3 Les axiomes

Les axiomes sont des assertions vraies concernant les abstractions du domaine [51].

D.3.4 Les instances

Les instances représentent la définition de l'extension de l'ontologie en véhiculant les connaissances du domaine [51].

D.4 Définition formelle

On peut définir une ontologie de la façon suivante en s'appuyant sur les travaux de Stumme & al. 2003 [52].

Soit une structure $C := (C, <_C, R, <_R, \sigma)$ **Avec :**

- C : Ensemble des identifiants de concepts
- R : Ensemble des identifiants de relations
- $<_C$: Ordre partiel sur C (hiérarchie de concepts)
- $<_R$: Ordre partiel sur R (hiérarchie de relation)
- Signature : $\sigma : R \to C^+$ qui définit le nombre d'instances de concepts liées par la relation, le type et l'ordre des concepts
- La définition mathématique de l'extension des concepts [c] et des relations [r]
- Un Système des axiomes

Exemple [32]:
Le lien relationnel entre les concepts « Texte » et « Auteur » est « A pour auteur ».

D.5 Les taxonomies

Selon la définition d'origine, elle se réfère au domaine de la classification des espèces :
« La taxonomie 'ou taxinomie' est la science qui décrit les organismes vivants et les regroupe en entités appelées « Taxons » en vue de les identifier, les nommer puis les classer ».Wikipédia
En fait, c'est une mesure de la distance évolutionnaire entre les espèces.

Dans le domaine informatique, le terme « Taxonomie » définie une méthode de classification de données dans une forme d'architecture évolutive structurée.
Dans le domaine des mathématiques, une taxonomie hiérarchique est une structure arborescente de classifications pour un ensemble d'objets. La racine de cette arborescence représente une classification unique s'appliquant à tout le reste des objets. Plus un nœud est bas dans l'arborescence, plus il représente une classification plus spécifique.

Exemple :
Voici le système taxonomique de Carolus Linnaeus[14] utilisé par les taxonomistes.

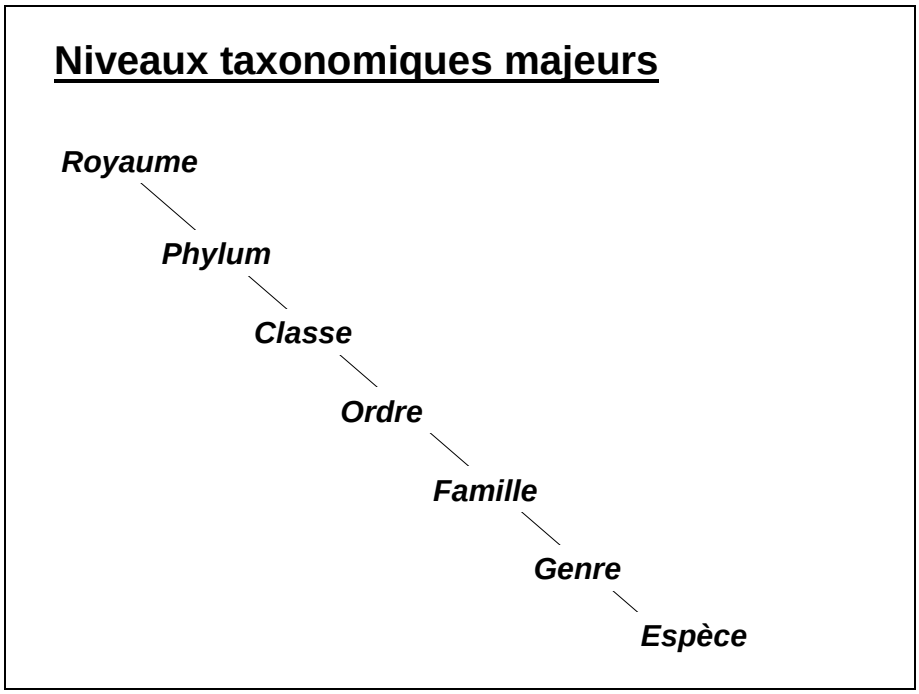


Figure 26: Système de Linnaeus décomposant les organismes en 7 divisions majeures

[14] **Carl von Linné** (après son anoblissement) (1707, 1778) est un naturaliste suédois père de la taxonomie moderne.

ANNEXE E

XBRL

E.1 Définition

XBRL est une méthode d'étiquetage des données en vue d'obtenir un format interprétable par un grand nombre de logiciels [53]. XBRL est basé sur la syntaxe XML à travers laquelle on associe chaque information à ses métadonnées (description contenant la signification et le contexte) [54].
Pour partager de l'information financière, toute organisation peut y insérer des étiquettes reconnues par XBRL. Ces informations brutes sont souvent disséminées dans des publications web tels les rapports annuels des entreprises. Aussi, il est plus facile de comparer les entreprises entre elles pour une meilleure prise de décision.

E.2 Utilisation

Touchant l'information financière de façon générale, XBRL intéresse de plus en plus les organismes gouvernementaux de réglementation, les marchés des capitaux ainsi que les services financiers. Il a en outre l'appui des ordres professionnels et des normalisateurs comptables [53].

Suivant le même chemin que les projets d'élaboration de taxonomies respectant les PCGR (Principes Comptables Généralement Reconnus) américains et les normes comptables internationales, XBRL Canada est actuellement à l'étape finale d'élaboration d'une taxonomie canadienne (états financiers conçus selon le PCGR).
XBRL Canada fait équipe avec le CSA (*Canadian Securities Administrators*) qui est en cours d'établissement d'un programme qui va permettre aux émetteurs de volontairement déposer des états financiers en format XBRL sous SEDAR[1] [55].

E.3 Exemple

Information représentée [56]

CURRENT ASSETS	
Assets Held for Sale	100,000
Construction in Progress, Current	100,000
Inventories	100,000
Construction in Progress, Current	100,000
Hedging Instruments, Current [Asset]	100,000
Current Tax Receivables	100,000
Trade and Other Receivables, Net, Current	100,000
Prepayments, Current	100,000
Cash and Cash Equivalents	100,000
Other Assets, Current	100,000
Current assets, Total	1,000,000

Figure 27 Information résultante d'une représentation XBRL

Format XBRL

```
<ifrs-gp:AssetsHeldSale contextRef="Current_AsOf" unitRef="U-Euros"
  decimals="0">100000</ifrs-gp:AssetsHeldSale>
<ifrs-gp:ConstructionProgressCurrent contextRef="Current_AsOf"
  unitRef="U-Euros" decimals="0">100000</ifrs-
  gp:ConstructionProgressCurrent>
<ifrs-gp:Inventories contextRef="Current_AsOf" unitRef="U-Euros"
  decimals="0">100000</ifrs-gp:Inventories>
<ifrs-gp:OtherFinancialAssetsCurrent contextRef="Current_AsOf"
  unitRef="U-Euros" decimals="0">100000</ifrs-
  gp:OtherFinancialAssetsCurrent>
<ifrs-gp:HedgingInstrumentsCurrentAsset contextRef="Current_AsOf"
  unitRef="U-Euros" decimals="0">100000</ifrs-
  gp:HedgingInstrumentsCurrentAsset>
<ifrs-gp:CurrentTaxReceivables contextRef="Current_AsOf" unitRef="U-
  Euros" decimals="0">100000</ifrs-gp:CurrentTaxReceivables>
<ifrs-gp:TradeOtherReceivablesNetCurrent contextRef="Current_AsOf"
  unitRef="U-Euros" decimals="0">100000</ifrs-
  gp:TradeOtherReceivablesNetCurrent>
<ifrs-gp:PrepaymentsCurrent contextRef="Current_AsOf" unitRef="U-Euros"
  decimals="0">100000</ifrs-gp:PrepaymentsCurrent>
<ifrs-gp:CashCashEquivalents contextRef="Current_AsOf" unitRef="U-
  Euros" decimals="0">100000</ifrs-gp:CashCashEquivalents>
<ifrs-gp:OtherAssetsCurrent contextRef="Current_AsOf" unitRef="U-Euros"
  decimals="0">100000</ifrs-gp:OtherAssetsCurrent>
<ifrs-gp:AssetsCurrentTotal contextRef="Current_AsOf" unitRef="U-Euros"
  decimals="0">1000000</ifrs-gp:AssetsCurrentTotal>
```

Figure 28: Expression d'une information sous le format XBRL

ANNEXE F

Liste des normes IFRS

Lien : http://fr.wikipedia.org/wiki/IFRS (site consulté le 20 juillet 2009)

Module	Name	Domaines d'analyse
IAS 1	Presentation of Financial Statements	Présentation des états financiers (voir bilan, compte de résultat, notes etc.)
IAS 2	Inventories	Stocks
IAS 7	Cash Flow Statements	Tableaux des flux de trésorerie
IAS 8	Accounting Policies, Changes in Accounting Estimates and Errors	Méthodes comptables, changements d'estimations et corrections d'erreurs
IAS 10	Events After the Balance Sheet Date	Événements postérieurs à la date de clôture
IAS 11	Construction Contracts	Contrats de construction
IAS 12	Income Taxes	Impôts sur le résultat
IAS 16	Property, Plant and Equipment	Immobilisations corporelles
IAS 17	Leases	Contrats de location (voir crédit-bail et immobilisation corporelle)
IAS 18	Revenue	Produits des activités ordinaires (voir chiffre d'affaires)
IAS 19	Employee Benefits	Avantages du personnel
IAS 20	Accounting for Government Grants and Disclosure of Government Assistance	Comptabilisation des subventions publiques et informations à fournir
IAS 21	The Effects of Changes in Foreign Exchange Rates	Effets des variations des cours des monnaies étrangères (voir taux de change, risque de change)
IAS 23	Borrowing Costs	Coûts d'emprunt
IAS 24	Related Party Disclosures	Information relative aux parties liées
IAS 26	Accounting and Reporting by Retirement Benefit Plans	Comptabilité et reporting des engagements de retraite
IAS 27	Consolidated and Separate Financial Statements	États financiers consolidés et individuels
IAS 28	Investments in Associates	Participations dans des entreprises associées (<50% du capital)
IAS 29	Financial Reporting in Hyperinflationary Economies : voir	Information financière dans les économies hyperinflationnistes
IAS 30	Disclosures in the Financial Statements of Banks and Similar Financial Institutions	Informations à fournir dans les états financiers des banques et des institutions financières assimilées (norme supprimée)
IAS 31	Interests in Joint Ventures	Participations dans des coentreprises
IAS 32	Financial Instruments (Disclosure and Presentation)	Instruments financiers - Présentation

Module	Name	Domaines d'analyse
IAS 33	Earnings per Share	Résultat par action
IAS 34	Interim Financial Reporting	Information financière intermédiaire
IAS 36	Impairment of assets	Dépréciation d'actifs (voir amortissement)
IAS 37	Provisions, Contingent Liabilities and Contingent Assets	Provisions, passifs éventuels et actifs éventuels
IAS 38	Intangible Assets	Immobilisations incorporelles
IAS 39	Financial Instruments (Recognition and Measurement)	Instruments financiers - Comptabilisation et évaluation
IAS 40	Investment Property	Immeubles de placement
IAS 41	Agriculture	Agriculture
IFRS 1	First-time Adoption of International Financial Reporting Standards	Première application des normes IFRS
IFRS 2	Share-based Payment	Paiement fondé sur des actions
IFRS 3	Business Combinations	Regroupements d'entreprises (voir fusions, acquisitions, goodwill)
IFRS 4	Insurance Contracts	Contrats d'assurance
IFRS 5	Non-current Assets Held for Sale and Discontinued Operations	Actifs non courants destinés à être vendus et activités abandonnées
IFRS 6	Exploration for and Evaluation of Mineral resources	Prospection et évaluation des ressources minérales
IFRS 7	Financial Instruments: Disclosures	Instruments financiers - Information à fournir
IFRS 8	Operating segments	Secteurs opérationnels

Table 31: Liste des normes IFRS

ANNEXE G

Notes IFRS

Les notes suivantes sont afférentes aux états financiers concernant les regroupements d'entreprises.

G.1 Liste des notes IFRS

[817000] Notes - Business combinations		
Disclosure of business combinations	explanatory text	IFRS 3 - Disclosures, IFRS 3 - Disclosures [2007-03-01]
Description of nature and financial effect of business combinations during period	text	IFRS 3.59 (a) , IFRS 3.66 (a) [2007-03-01]
Description of nature and financial effect of business combinations after reporting period before statements authorised for issue	text	IFRS 3.66 (b) [2007-03-01] , IFRS 3.59 (b)
Explanation of financial effect of adjustments related to business combinations	text	IFRS 3.61
Disclosure of information for each business combination		IFRS 3.B67, IFRS 3.B64
Name of acquiree	text	IFRS 3.B64 (a)
Description of acquiree	text	IFRS 3.B64 (a)
Date of acquisition	yyyy-mm-dd	IFRS 3.B64 (b)
Percentage of voting equity interests acquired	X.XX	IFRS 3.B64 (c)
Description of primary reasons for business combination	text	IFRS 3.B64 (c)
Description of how acquirer obtained control of acquiree	text	IFRS 3.B64 (d)
Description of factors that make up goodwill recognised	text	IFRS 3.B64 (e)
Acquisition-date fair value of total consideration transferred	X	IFRS 3.B64 (f)
Cash transferred	X	IFRS 3.B64 (f)
Other tangible or intangible assets transferred	X	IFRS 3.B64 (f)
Liabilities incurred	X	IFRS 3.B64 (f)
Equity interests of acquirer	X	IFRS 3.B64 (f)
Number of instruments or interests issued or issuable	X.XX	IFRS 3.B64 (f)
Method of determining fair value of instruments or interests	text	IFRS 3.B64 (f)
Contingent consideration arrangements and indemnification assets recognised as of acquisition date	X	IFRS 3.B64 (g)
Description of arrangement for contingent consideration arrangements and indemnification assets	text	IFRS 3.B64 (g)
Description of basis for determining amount of payment for contingent consideration arrangements and indemnification assets	text	IFRS 3.B64 (g)
Description of estimate of range of outcomes from contingent consideration arrangements and indemnification assets	text	IFRS 3.B64 (g)
Description of explanation of fact and reasons why range of outcomes from contingent consideration arrangements and indemnification assets cannot be estimated	text	IFRS 3.B64 (g)
Explanation of fact that maximum amount of payment for contingent consideration arrangements and indemnification assets is unlimited	text	IFRS 3.B64 (g)
Fair value of acquired receivables	X	IFRS 3.B64 (h)
Gross contractual amounts receivable for acquired receivables	X	IFRS 3.B64 (h)
Explanation of best estimate at acquisition date of contractual cash flows not expected to be collected for acquired receivables	text	IFRS 3.B64 (h)
Description of amounts recognised as of acquisition date for each major class of assets acquired and liabilities assumed	text	IFRS 3.B64 (i)
Information required in paragraph 85 of IAS 37 for each contingent liability recognised	text	IFRS 3.B64 (j)
Information required by paragraph 86 of IAS 37 if contingent liability is not recognised because fair value cannot be measured reliably	text	IFRS 3.B64 (j)

Description of reasons why liability cannot be measured reliably	text		IFRS 3.B64 (j)
Goodwill expected deductible for tax purposes		X	IFRS 3.B64 (k)
Explanation of transactions recognised separately from acquisition of assets and assumption of liabilities in business combination	text		IFRS 3.B64 (m) , IFRS 3.B64 (l)
Explanation of amount of any gain recognised and line item in statement of comprehensive income in which gain is recognised in bargain purchase	text		IFRS 3.B64 (n)
Description of reasons why transaction resulted in gain in bargain purchase	text		IFRS 3.B64 (n)
Non-controlling interest in acquiree recognised at acquisition date		X	IFRS 3.B64 (o)
Description of measurement basis for non-controlling interest in acquiree recognised at acquisition date	text		IFRS 3.B64 (o)
Description of valuation techniques and key model inputs used for determining non-controlling interest in an acquiree measured at fair value	text		IFRS 3.B64 (o)
Acquisition-date fair value of equity interest in acquiree held by acquirer immediately before acquisition date		X	IFRS 3.B64 (p)
Description of amount of any gain or loss recognised as result of remeasuring to fair value equity interest in acquiree held by acquirer before business combination and line item in statement of comprehensive income in which that gain or loss is recognised	text		IFRS 3.B64 (p)
Revenue of acquiree		X	IFRS 3.B64 (q)
Profit (loss) of acquiree		X	IFRS 3.B64 (q)
Revenue of combined entity		X	IFRS 3.B64 (q)
Profit (loss) of combined entity		X	IFRS 3.B64 (q)
Description of reasons why initial accounting for business combination is incomplete	text		IFRS 3.B67 (a)
Description of amounts of assets, liabilities, equity interests or items of consideration for which initial accounting is incomplete	text		IFRS 3.B67 (a)
Description of nature and amount of any measurement period adjustments recognised for particular assets, liabilities, non-controlling interests or items of consideration	text		IFRS 3.B67 (a)
Explanation of any changes in recognised amounts of contingent consideration	text		IFRS 3.B67 (b)
Explanation of any changes in range of outcomes (undiscounted) and reasons for those changes for contingent consideration	text		IFRS 3.B67 (b)
Description of valuation techniques and key model inputs used to measure contingent consideration	text		IFRS 3.B67 (b)
Disclosure of information required by paragraphs 84 and 85 of IAS 37 for each class of provision	text		IFRS 3.B67 (c)
Reconciliation of carrying amount of goodwill			IFRS 3.B67 (d)
Gross amount of goodwill at beginning of period		X	IFRS 3.B67 (d) , IFRS 3.B67 (d)
Accumulated impairment losses of goodwill at beginning of period		X	IFRS 3.B67 (d) , IFRS 3.B67 (d)
Additional goodwill recognised		X	IFRS 3.B67 (d)
Increase (decrease) of goodwill resulting from subsequent recognition of deferred tax assets		X	IFRS 3.B67 (d)
Goodwill included in disposal group classified as held for sale		X	IFRS 3.B67 (d)
Goodwill derecognised without having previously been included in disposal group classified as held for sale		X	IFRS 3.B67 (d)
Impairment losses of goodwill recognised in accordance with IAS 36		X	IFRS 3.B67 (d)

Net exchange rate differences of goodwill	X	IFRS 3.B67 (d)
Other changes in carrying amount of goodwill	X	IFRS 3.B67 (d)
Gross amount of goodwill at end of period	X	IFRS 3.B67 (d) , IFRS 3.B67 (d)
Accumulated impairment losses of goodwill at end of period	X	IFRS 3.B67 (d) , IFRS 3.B67 (d)
Gain (loss) that relates to identifiable assets acquired or liabilities assumed in business combination	X	IFRS 3.B67 (e)
Explanation of gain or loss that relates to identifiable assets acquired or liabilities assumed in business combination	text	IFRS 3.B67 (e)
Gain (loss) that is of such size, nature or incidence that disclosure is relevant to understanding combined entity's financial statements	X	IFRS 3.B67 (e)
Explanation of gain or loss that is of such size, nature or incidence that disclosure is relevant to understanding combined entity's financial statements	text	IFRS 3.B67 (e)
Distinction of each business combination		IFRS 3.B64
Business combinations	for each	IFRS 3.B64
Business combinations		IFRS 3.B64
Aggregated individually immaterial business combinations		IFRS 3.B65
Disclosure of fact and explanation why disclosure of information for each business combination is impracticable	text	IFRS 3.B64
Explanation which disclosures could not be made and reasons why they cannot be made if initial accounting for business combination is incomplete	text	IFRS 3.B66
Disclosure of business combinations and goodwill	explanatory text	IFRS 3 - Disclosures [2007-03-01]
Disclosure of business combinations [explanatory]	explanatory text	IFRS 3 - Disclosures, IFRS 3 - Disclosures [2007-03-01]
Description of nature and financial effect of business combinations during period	text	IFRS 3.59 (a) , IFRS 3.66 (a) [2007-03-01]
Description of nature and financial effect of business combinations after reporting period before statements authorised for issue	text	IFRS 3.66 (b) [2007-03-01] , IFRS 3.59 (b)
Explanation of financial effect of gain or loss related to business combinations	text	IFRS 3.72 [2007-03-01]
Explanation of financial effect of error corrections related to business combinations	text	IFRS 3.72 [2007-03-01]
Explanation of financial effect of other adjustments related to business combinations	text	IFRS 3.72 [2007-03-01]
Disclosure of goodwill	explanatory text	IFRS 3 - Disclosures [2007-03-01]
Additional information about changes during period	text	IFRS 3.74 [2007-03-01]
Explanation of goodwill included in disposal group classified as held for sale	text	IFRS 3.75 (d) [2007-03-01]
Explanation of recognition and derecognition of goodwill during period	text	IFRS 3.74 [2007-03-01]

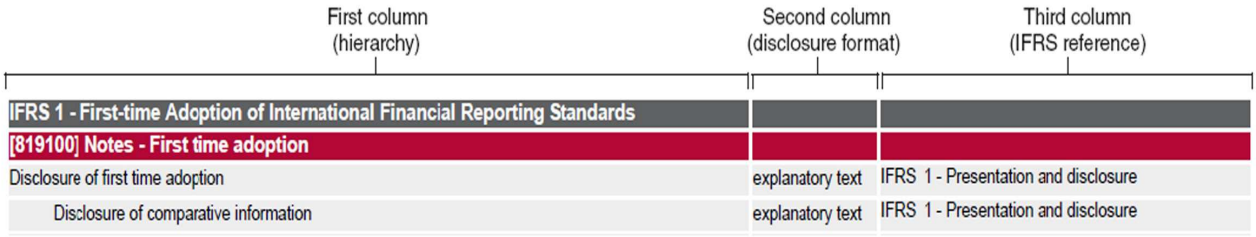
Source : http://eifrs.iasb.org/eifrs/Taxonomy?type=r&lang=en

G.2 Source de l'information

Pour consulter la source d'information contenant la liste des notes IFRS, se reporter au site web:
http://eifrs.iasb.org/eifrs/taxonomy/guide.html

G.3 Lecture de la taxonomie IFRS illustrée

Cette section décrit le format et le contenu de la taxonomie IFRS illustrée. Les explications données ci-dessous s'appliquent à l'ensemble du document.



G.3.1 Première colonne – Hiérarchie

La première colonne du document représente l'hiérarchie de la taxonomie IFRS:

- Les rubriques de la colonne représentent un IFRS, un IAS ou une interprétation
- Les sous-rubriques de la colonne représentent le nom d'un composant taxonomique IFRS. Chaque sous-rubrique de colonne est précédée par un numéro à six chiffres entre crochets avec une valeur comprise entre [100000] et [999999]. Ces numéros sont artificiels et offrent l'affichage et la fonctionnalité de tri (Ils ne sont pas liés aux IFRS)
- Les lignes en dessous des sous-rubriques de colonne représentent les éléments appartenant à ce composant, et qui sont les exigences de la norme IFRS.

G.3.2 Deuxième colonne –format de divulgation

La seconde colonne du document illustre les formats possibles qu'une divulgation donnée peut prendre. Ceux-ci sont les suivants:

- Texte explicatif- dénote que le format de divulgation est un texte explicatif
- Texte – dénote que le format de divulgation est un texte
- yyyy-mm-dd – dénote que le format de divulgation est une date
- X – dénote que le format de divulgation est une valeur monétaire
- (X) – dénote que le format de divulgation est une valeur monétaire présentée comme étant négative
- X.XX – dénote que le format de divulgation est une valeur décimale
- shares – dénote que le format de divulgation est un nombre de parts
- ____ dénote que le format de divulgation est la somme totale des lignes précédentes
- For each – dénote une divulgation répétitive dans laquelle les éléments des lignes précédentes sont révélés pour chacun des éléments des lignes subséquentes
- Une colonne vide dénote qu'une divulgation n'est pas nécessaire

G.3.3 Troisième colonne – référence IFRS

La troisième colonne indique les IFRS paragraphe/section correspondants à une divulgation donnée. Sauf indication contraire, la date de ces documents est fixée au 1 er Janvier 2009 (relativement au volume relié des IFRS et des déclarations officielles publiées au 1er Janvier 2009). Dans la version PDF électronique, ces références contiennent des liens vers les IFRS électroniques (eIFRS).

ANNEXE H

Cycle de vie du projet

H.1 Vue générale

En résumé, après avoir présenté les concepts qui sous-tendent notre recherche, nous rappelons dans cette courte section, sous forme schématique (voir la Figure 29), les grandes étapes techniques à réaliser dans ce projet :

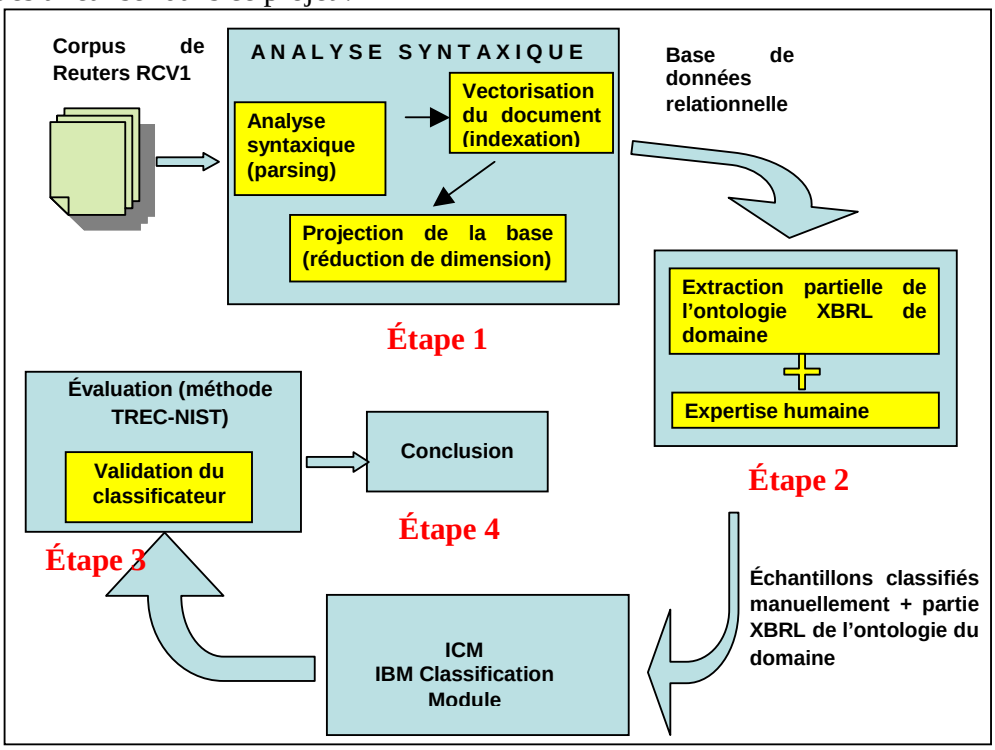


Figure 29: Cycle de vie du projet de recherche

H.2 Organigramme schématisé des opérations hiérarchiques effectuées

Soit un organigramme schématisé représentant toutes les opérations techniques faites ou à faire dans le présent projet (voir la Figure 30).

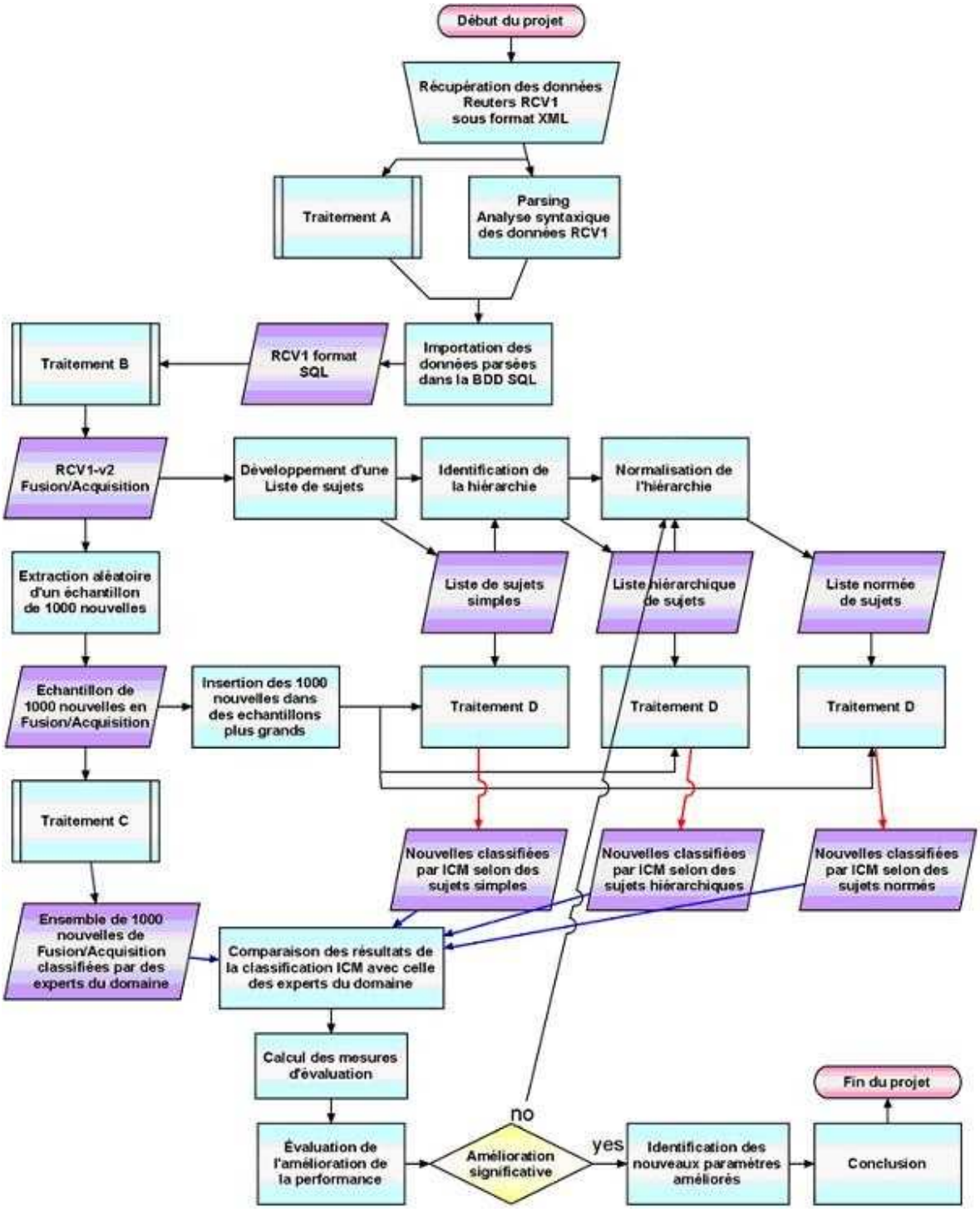


Figure 30: Organigramme principal du projet

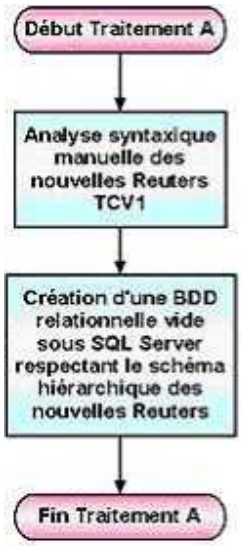


Figure 31: Organigramme partiel représentant l'analyse syntaxique de RCV1

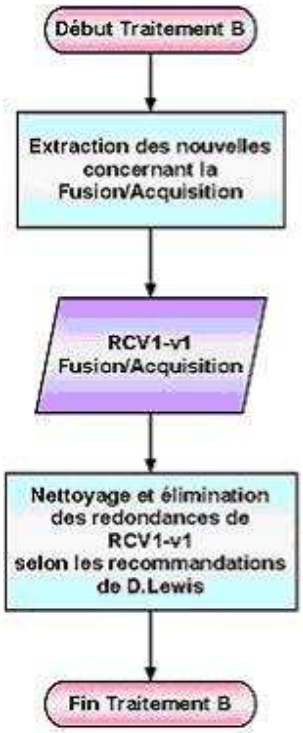


Figure 32: Organigramme partiel d'extraction et d'épuration des nouvelles

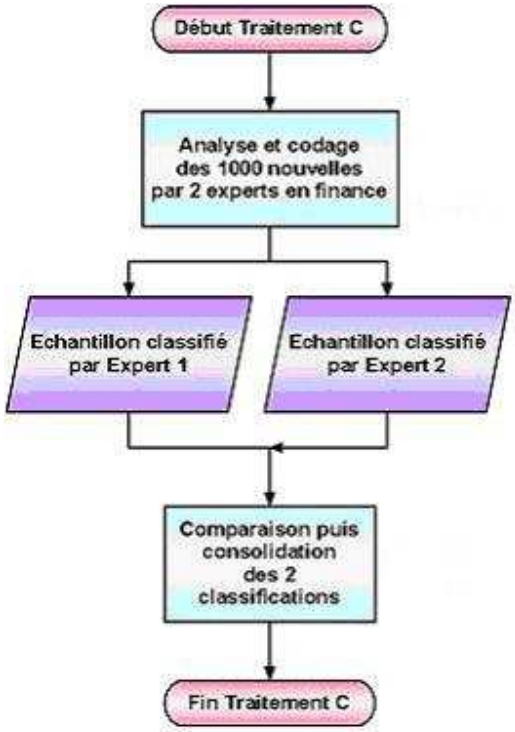


Figure 33: Organigramme partiel de classification manuelle d'un échantillon de nouvelles

ANNEXE I

Corpus RCV1 de Reuters

I.1 Introduction

Reuters est une collection de nouvelles disponible, depuis l'automne 2004. Le NIST se charge de sa distribution. Les données de la collection ont été formatées sous XML et sont organisées autour de trois ensembles de catégories principales : les thèmes, les industries, et les régions. Nous utilisons seulement les 41 214 nouvelles liées au code C181, *Mergers and Acquisitions* (*fusion* et *acquisition*), appartenant au code C18, *Ownership Changes*. Le nombre de nouvelles a été réduit pour nettoyer la base des nouvelles incomplètes et produire le RCV1 version 2 [18]. Cette collection nécessite néanmoins un pré traitement avant son utilisation afin d'extraire les données importantes identifiant le type de chaque nouvelle. Le type est contenu dans un code spécifiant, entre autres, si la nouvelle concerne une *fusion*/*acquisition* quelconque.

I.2 Correction des faiblesses du RCV1 et épuration des nouvelles extraites

« *Apart from the terrible memories this stirs up*
for me personally (coding stories through the
night etc.), I can't find fault with your account. »

– Éditeur Reuters commentant le travail de David Lewis.

David Lewis a procédé à la correction de la 1ère version de RCV1 après avoir étudié ses codes et sa sémantique. Pourtant, la version RCV1 était beaucoup moins bruitée et contient beaucoup moins d'erreurs que les collections Reuters qui l'ont précédée. Dans RCV1, un identificateur unique était affecté à chaque document, le format XML du texte et des métadonnées facilitait son utilisation et il contient la plupart des nouvelles d'un type particulier étalées sur un intervalle d'une année. Cela n'a pourtant pas empêché que certaines erreurs, dues au travail en batch long et ardu, se sont glissées. Pour cette raison, David Lewis a rédigé un nouveau corpus RCV1 version 2 avec un ensemble de corrections.

Sachant que les données de la collection RCV1 sont organisées autour de trois ensembles de catégories principales (sujets, industries et régions), David Lewis a noté une liste d'erreurs parmi lesquelles on note les plus importantes [19]:

1. Le nombre des codes de sujets (*topics*) existants est 126 alors que la collection utilise 103.
2. Le nombre des codes industries existants est 870 alors que la collection utilise 354.
3. Le nombre des codes de régions existants est 366 alors que la collection utilise 296. En plus de l'existence de 4 codes dans la collection qui ne font pas partie de la liste des catégories des régions.
4. Certaines nouvelles sont dupliquées faussant les calculs de probabilité d'apparition et rendant les résultats de classification erronés.
5. Les anomalies dans le codage hiérarchique des industries.

David Lewis a fourni une solution à toutes ces erreurs ainsi qu'un nouveau codage et a permis de développer une nouvelle collection beaucoup moins bruitée et plus intéressante d'utilisation pour le domaine des catégorisations.

Dans le cas de notre projet, les nouvelles ont été traitées selon le contenu de leurs métadonnées implicites et non selon les catégories fournies par Reuters. Le cas des codes en trop ne se pose donc pas dans notre étude (voir la Figure 37). Par contre, une liste de documents dupliqués a été détectée grâce à la similitude de leur « *Headline* » et de leur « *Text* ». Pour cette raison, une petite application de nettoyage a été développée sous le Visual Basic qui nous a permis d'obtenir un ensemble de nouvelles financières en *Fusion/Acquisition* non dupliquées propres à une utilisation non bruitée dans la suite du processus de ce projet d'étude.

I.3 Analyse syntaxique des données RCV1

L'analyse syntaxique des données RCV1 repose sur une application en VB, développée dans le cadre de ce projet, traitant séquentiellement chacune des 810 000 nouvelles pour les transformer en fichiers relationnels contenant les données importantes des nouvelles avant de les exporter sous le format d'une base de données relationnelle sous SQL Server 2005 (*Structured Query Language Server*). Nous obtenons alors le schéma relationnel de la Figure 34 qui représente les différents liens relationnels entre les tables de notre base de connaissance issue de RCV1. La description de chaque table est décrite ci bas :

- la table « ***News*** » contient la liste des nouvelles et leurs métadonnées.
- la table « ***Paragraphes*** » contient tous les textes des nouvelles qui seront analysés par la suite par le classificateur en utilisant la NLP.
- la table « ***Codes*** » contient une liste de codes représentant, selon le cas, un code d'industrie, de région ou de sujet (*Topic*) que l'on retrouve regroupés dans le reste des tables (« ***Industries*** », « ***Regions*** » et « ***Topics*** »).
- La table « ***Topics*** » va nous permettre d'extraire les données touchant les sujets de fusion/acquisition.

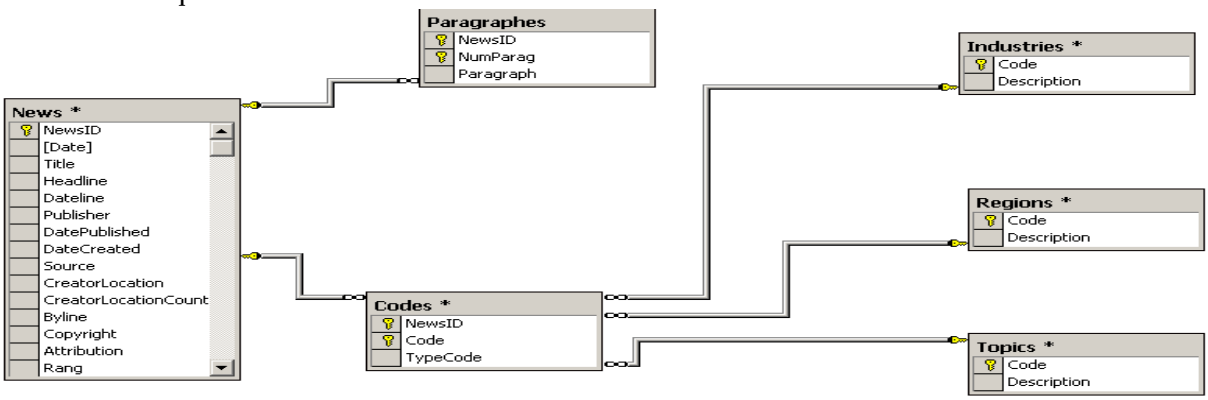


Figure 34: Schéma relationnel de la base de données représentant RCV1

I.4 Extraction des nouvelles de *fusion/acquisition*

Grâce à la base de données relationnelle obtenue, nous extrayons toutes les nouvelles dont le code financier doit être égal à « C181 » et qui concerne la *fusion/acquisition* en utilisant des requêtes SQL.

Voici un exemple de requête utilisée :

```
SELECT Codes.NewsID, Date, Title, Headline,
DateLine,Publisher,DatePublished,DateCreated,Source,CreatorLocation,CreatorL
ocationCountryName,Byline,Copyright,Attribution INTO LN
FROM Codes, News
WHERE Codes.NewsID=News.NewsID AND Codes.TypeCode='T' AND Codes.code='C181'
ORDER BY Date, Codes.NewsID
```

Figure 35: Exemple d'une requête SQL pour l'extraction de nouvelles

I.5 Extraction aléatoire d'un échantillon de 1000 nouvelles

Grâce à un ensemble d'applications, un nombre de nouvelles aléatoires, basées sur la fusion/acquisition, est extrait en vue de le traiter dans les prochains processus. L'exemple suivant échantillonne 41 214 nouvelles en *fusion* et *acquisition*:

```
Sub EchantillonnageDeMillesNouvelles()
' EchantillonnageDeMillesNouvelles Macro
' Macro enregistrée le 2009-07-13 par Sadia Messaoudi
Dim counter, myNum, myValue
counter = 41214, j = 1
   Do While j <= 1000
      myValue = Int((counter * Rnd) + 1)
      Fichier = Worksheets(1).Cells(myValue, 1).Value
      Worksheets(2).Cells(j, 1).Value = Fichier
      Worksheets(1).Cells(myValue, 1).Delete
      j = j + 1, counter = counter - 1
   Loop
   MsgBox "Echantillonnage fini."
End Sub
```

Figure 36: Exemple d'un programme d'échantillonnage en Visual Basic

I.6 Exemple d'une nouvelle Reuters

```
<?xml version="1.0" encoding="iso-8859-1" ?>
- <newsitem itemid="760540" id="root" date="1997-07-28" xml:lang="en">
  <title>USA: FULL TEXT - IMC Global in deal with Freeport.</title>
  <headline>FULL TEXT - IMC Global in deal with Freeport.</headline>
  <dateline>NEW YORK 1997-07-28</dateline>
- <text>  <p>IMC Global Inc and Freeport-McMoran Inc signed a letter of intent
to merge IGL and FTX in a stock transaction. IGL will be the surviving entity. The
sulphur business and the 58.3 percent interest in the Main Pass Block 299
sulphur and oil and gas operations owned by Freeport-McMoRan Resource
Partners, Limited Partnership and the 25 percent interest in Main Pass 299
```

```
owned by IGL, will be transferred to a newly-formed subsidiary of FRP. The unit will distribute the Newco shares to all FRP unitholders, including FTX.</p>
  <p>The merger is expected to result in an annual general and administrative cash cost savings of at least $33 million immediately from the elimination of FTX and FRP costs. This savings amount is expected to increase to approximately $40 million per year in the next several years. In addition, IGL expects additional opportunities for further cost savings from this combination.</p> </text>
  <copyright>(c) Reuters Limited 1997</copyright>
- <metadata>
- <codes class="bip:topics:1.0">
- <code code="C18">
    <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1997-07-28" /> </code>
- <code code="C181">
    <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1997-07-28" /> </code>
- <code code="CCAT">
    <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1997-07-28" /> </code>
  </codes>
  <dc element="dc.date.created" value="1997-07-28" />
  <dc element="dc.publisher" value="Reuters Holdings Plc" />
  <dc element="dc.date.published" value="1997-07-28" />
  <dc element="dc.source" value="Reuters" />
  <dc element="dc.creator.location" value="NEW YORK" />
  <dc element="dc.creator.location.country.name" value="USA" />
  <dc element="dc.source" value="Reuters" />
  </metadata>
  </newsitem>
```


Figure 37: Exemple d'une nouvelle Reuters contenant des métadonnées clés

Dans l'exemple ci-haut, on constate les éléments les plus importants suivants représentant les détails et l'utilité de chaque nouvelle :

- ***itemID*** représente un identifiant unique de la nouvelle
- ***title*** contient le titre de la nouvelle
- ***text*** contient le corps de l'information véhiculée par la nouvelle. On retrouve ce champ représenté par les enregistrements de la table *Paragraphes* de notre base de données relationnelle (voir la section I.3). Cet élément est important puisqu'il va servir à l'analyse NLP d'ICM lors de la classification de la nouvelle.
- ***Code*** contient le code de la nouvelle. Il nous permet de filtrer les nouvelles pour n'extraire que celles concernant la *fusion/acquisition*.

ANNEXE J

Classification manuelle des experts

J.1 Introduction

La classification manuelle des nouvelles a été réalisée grâce à l'utilisation d'une application développée dans ce but. Cette application Access permet d'utiliser la liste normée dans le choix des classes pour les 1000 nouvelles de fusion/acquisition pré-extraites auparavant.
Les 2 experts du domaine l'utilisent chacun de façon indépendante de l'autre sur une copie des 1000 nouvelles afin de pouvoir obtenir 2 classifications différentes dans le but de la comparaison.
Un troisième expert du domaine utilisera l'application avec un nombre limité de nouvelles (moins d'une centaine) qui ont été classées par les 2 experts précédents de façon trop contradictoire. Cette contradiction a été déduite à travers les résultats obtenus par la mesure proposée. Les nouvelles candidates sont celles ayant enregistré une contradiction totale par rapport aux résultats de la classification automatique d'ICM versus les classifications manuelles des 2 experts précédents.

J.2 Opérations possibles proposées par l'application

Dans le menu général se trouvent toutes les opérations à effectuer sur l'échantillon de 1000 nouvelles (ou n'importe quel autre échantillon).

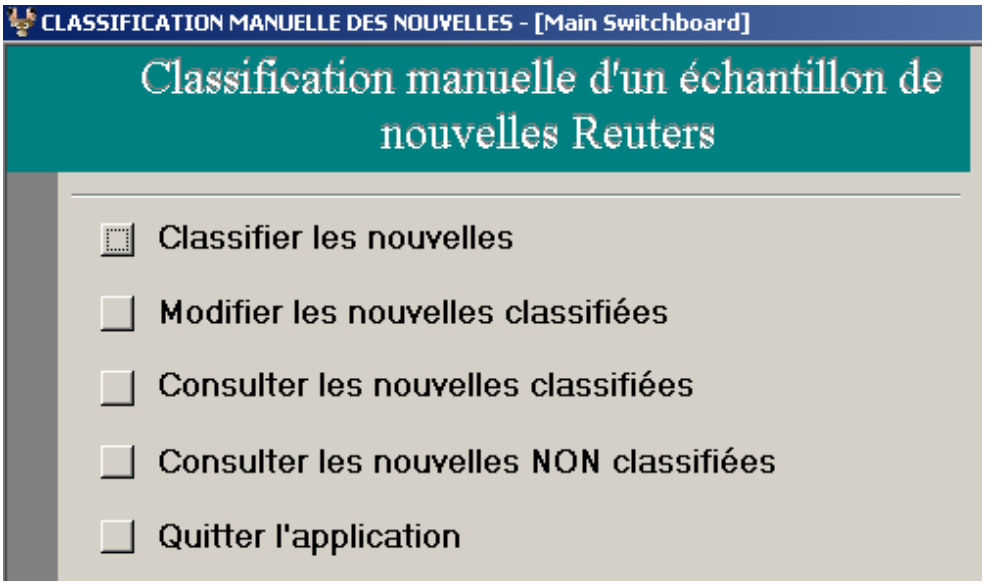


Figure 38 Menu général

Ainsi on peut effectuer les tâches suivantes :

J.2.1 Classifier les nouvelles

Dans ce menu, on effectue l'opération la plus importante de l'application : Classifier les nouvelles non encore classifiées (voir Figure 39):

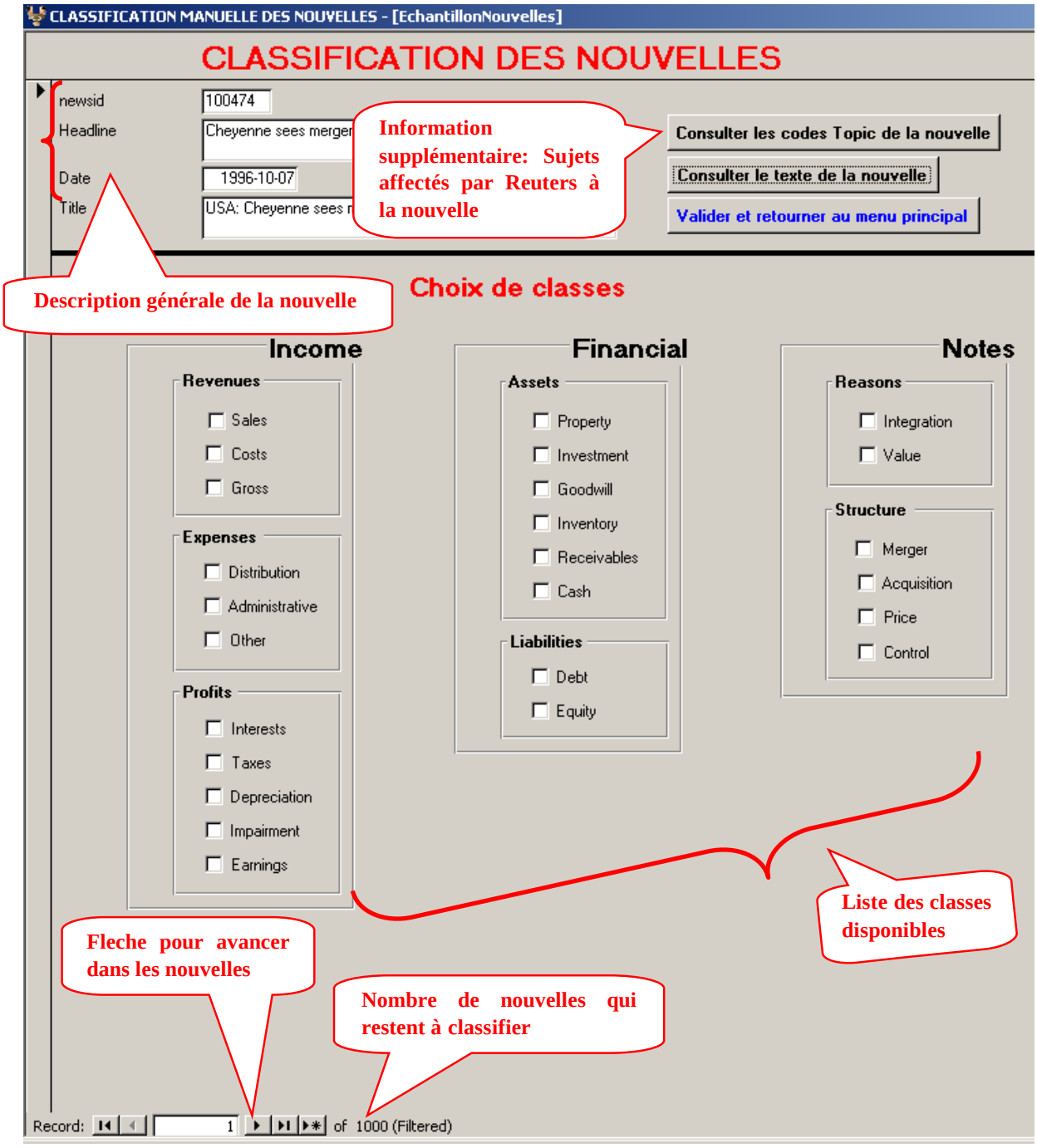


Figure 39 Formulaire de saisie des classes

Une liste de cases à cocher est offerte représentant la liste des classes disponibles parmi lesquelles on doit choisir une ou plusieurs classes pour chaque nouvelle. Cette liste contient 3 types de classes principales : ***Income***, ***Financial*** et ***Notes***.

J.2.2 Consulter le texte de la nouvelle

Pour consulter les lignes de paragraphe de la nouvelle en cours de traitement, on utilise le formulaire proposé par la 1ère option de consultation. Voir la Figure 40.

Dans l'exemple de la Figure 40, on a 14 lignes de textes qui parlent de la nouvelle en cours de traitement.

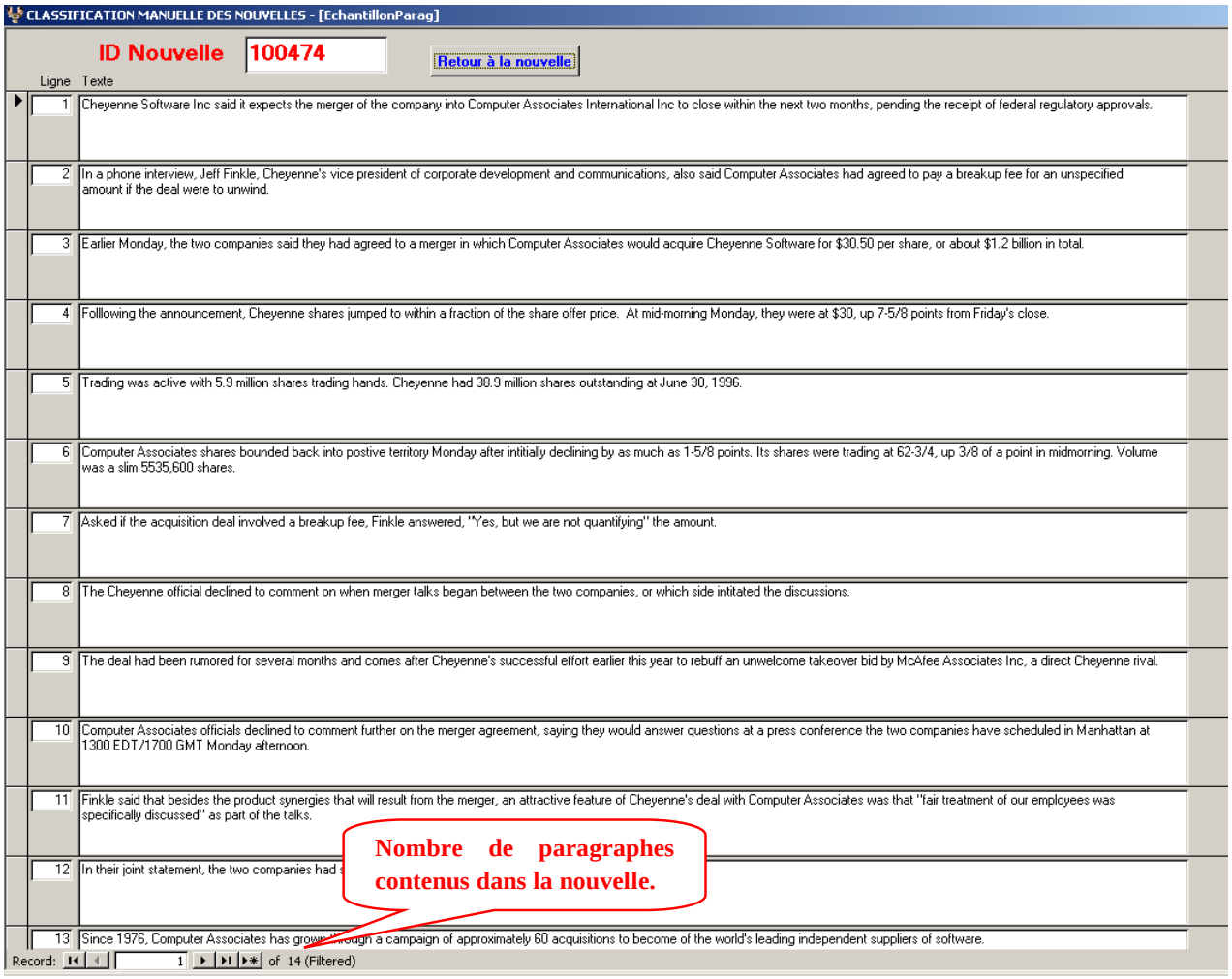


Figure 40 Texte de la nouvelle 100474

J.2.3 Consulter les codes *Topic* de la nouvelle

Pour consulter les sujets de la nouvelle affectés par Reuters, on utilise le formulaire proposé par la 2ème option de consultation. Voir la Figure 41.

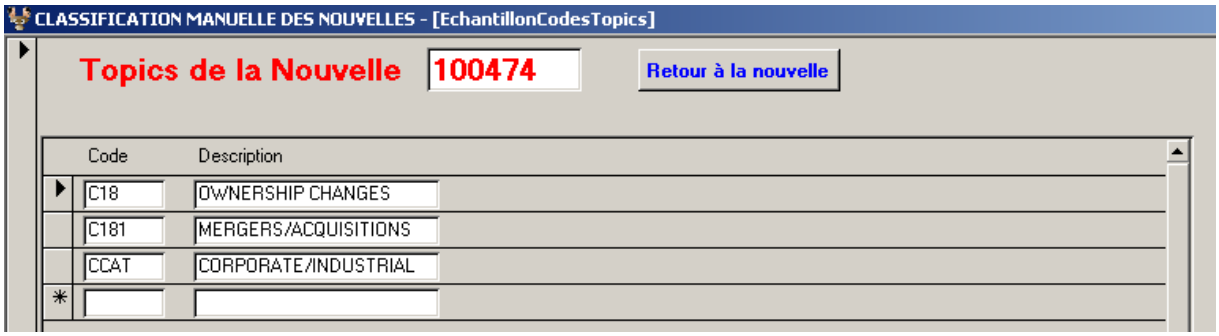


Figure 41 Liste des sujets de la nouvelle 100474

J.2.4 Modifier les nouvelles classifiées

On peut modifier à tout moment les nouvelles précédemment classifiées à travers ce menu se trouvant dans le menu général (voir Figure 39).

Le formulaire à utiliser est similaire au formulaire de saisie des classes à la seule différence que les seules nouvelles qu'on peut modifier sont celles qui ont déjà été classifiées.

J.2.5 Consulter les nouvelles classifiées

Grâce à ce menu, on peut consulter toutes les nouvelles classifiées sans pouvoir les modifier. En utilisant la 1ère option **[consulter les nouvelles classifiées]**, on obtient un sous-menu qui permet de consulter les nouvelles classifiées de 2 façons différentes (voir Figure 42) :

- L'option **[Nouvelles Classifiées – Consultation par liste]** permet de consulter la liste des nouvelles classifiées sans les détails des classes, des textes et des topics (voir Figure 43).

- L'option **[Nouvelles Classifiées – Consultation détaillée]** permet de consulter toutes les nouvelles classifiées avec les détails des classes choisies (voir Figure 42).

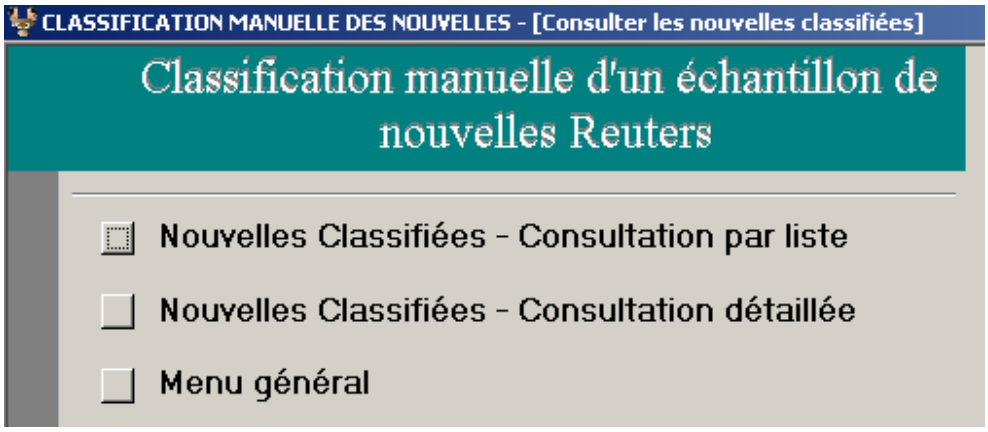


Figure 42 Sous-menu pour la consultation des nouvelles classifiées

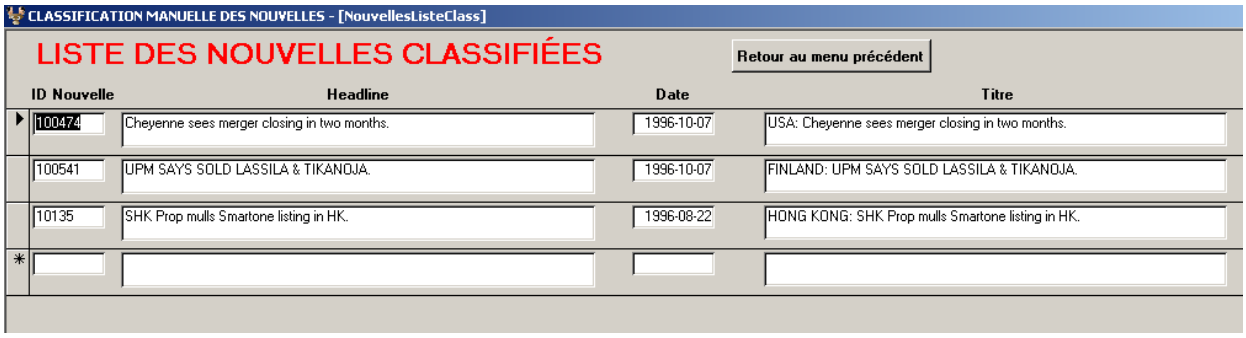


Figure 43 Liste des nouvelles classifiées en consultation uniquement

ANNEXE K

Résultats détaillés de la classification automatique des nouvelles

K.1 Classification des nouvelles sur la base des 14 classes dominantes

Comme la taille de l'ensemble des classes à utiliser à été fixé à 14, un ensemble de plus de 600 nouvelles a été extrait. Parmi cet ensemble, nous avons choisi 201 nouvelles pour l'entraînement et 201 autres nouvelles pour l'échantillon (voir la section 4.5.3 pour plus de détails). Nous avons obtenu des résultats de classification dont les performances sont calculées dans les sections suivantes grâce aux différentes mesures de performance présentées au chapitre précédent.

K.1.1 Résultats selon les mesures classiques de Sébastiani

EXPERT 1	Liste Simple	Liste Hiérarchique	Liste Normée
Précision	0,5230	0,6432	0,6385
Rappel	0,7886	0,8132	0,6678
F-Mesure	**0,6289**	**0,7183**	**0,6529**

EXPERT 2	Liste Simple	Liste Hiérarchique	Liste Normée
Précision	0,6228	0,7479	0,7341
Rappel	0,6569	0,6963	0,5550
F-Mesure	**0,6394**	**0,7211**	**0,6321**

Table 32: Résultats des mesures de Sébastiani sur la base de 201 nouvelles

Constat : La mesure considérée étant la F-Mesure qui permet d'équilibrer le poids du rappel et de la précision, nous avons constaté ce qui suit pour chaque expert (afin de mieux comprendre le raisonnement ayant permit de déduire la F-Mesure, se référer à l'annexe C, à la section C.2):

1. La valeur enregistrée par la liste hiérarchique est plus importante que celle de la liste simple (exemple : 0,7183 par rapport à 0,6289 pour l'expert 1)
2. La valeur enregistrée par la liste normée est moins importante que celle de la liste hiérarchique (exemple : 0,6529 par rapport à 0,7183 pour l'expert 1)

Même si la liste normée a enregistré des améliorations entre la liste normée et la liste simple (surtout dans le cas de l'expert 1) elle ne permet pas encore d'améliorer les performances par rapport à une liste hiérarchique dans le cas des mesures classiques de Sébastiani.

K.1.2 Résultats selon les mesures élaborées et Kiritchenko

Liste simple			
Mesures globales		**Exp1**	**Exp2**
Macro-moyenne	Macro-précision	0,4130	0,4813
	Macro-Rappel	0,5705	0,4282
	Macro-F-Mesure	**0,4791**	**0,4532**
Micro-moyenne	Micro-précision	0,5230	0,6228
	Micro-Rappel	0,7886	0,6569
	Micro-F-Mesure	**0,6289**	**0,6394**
Kiritchenko	Micro-précision	0,5230	0,6228
	Micro-Rappel	0,7886	0,6569
	Micro-F-Mesure	**0,6289**	**0,6394**

Table 33: Liste simple - Mesures de Kiritchenko suite à la classification de 201 nouvelles

Liste hiérarchique			
Mesures globales		**Exp1**	**Exp2**
Macro-moyenne	Macro-précision	0,4309	0,5311
	Macro-Rappel	0,5105	0,3525
	Macro-F-Mesure	**0,4674**	**0,4237**
Micro-moyenne	Micro-précision	0,6433	0,7479
	Micro-Rappel	0,8132	0,6963
	Micro-F-Mesure	**0,7183**	**0,7211**
Kiritchenko	Micro-précision	0,7089	0,8003
	Micro-Rappel	0,7795	0,8146
	Micro-F-Mesure	**0,7425**	**0,8074**

Table 34: Liste hiérarchique - Mesures de Kiritchenko pour 201 nouvelles

Liste normée			
Mesures globales		**Exp1**	**Exp2**
Macro-moyenne	Macro-précision	0,4555	0,5337
	Macro-Rappel	0,4319	0,3631
	Macro-F-Mesure	**0,4434**	**0,4322**
Micro-moyenne	Micro-précision	0,6386	0,7341
	Micro-Rappel	0,6679	0,5550
	Micro-F-Mesure	**0,6529**	**0,6321**
Kiritchenko	Micro-précision	0,7884	0,8192
	Micro-Rappel	0,9362	0,8852
	Micro-F-Mesure	**0,8560**	**0,8509**

Table 35: Liste normée - Mesures de Kiritchenko pour 201 nouvelles

Constat : Dans ce cas de figure, nous avons utilisé les macro et micro moyennes ainsi que les mesures de Kiritchenko afin de présenter l'impact de la parenté entre les classes dans le cas d'une classification hiérarchique des nouvelles. Ainsi, nous notons que pour chacune des 3 listes (simple, hiérarchique et normée) et chaque expert, nous avons calculé :

4. La micro-moyenne afin d'évaluer les performances globales de notre système de classification sans égard au poids des classes (voir la section 3.5.2 pour plus de détails).
5. La macro-moyenne afin de faire un estimé moyen des performances globales de notre système de classification.
6. Les mesures de Kiritchenko afin d'analyser le comportement du classificateur lorsque la parenté des classes est prise en considération.

Étant donné que les valeurs obtenues sont complexes et puisque la F-Mesure est une base de raisonnement que nous avons adopté dans ce projet, nous avons résumé les mesures en ne gardant que les F-Mesures pour chaque liste et chaque expert (voir la Table 36 et la Table 37). Ces valeurs ont été ensuite utilisées pour tracer un graphe pour chaque expert (voir la Figure 44 et la Figure 45).

E X P E R T 1 :			
Mesures	***Liste Simple***	***Liste Hiérarchique***	***Liste Normée***
Macro-F-Mesure	***0,4791***	***0,4674***	***0,4434***
Micro-F-Mesure	***0,6289***	***0,7183***	***0,6529***
Kiritchenko-F-Mesure	***0,6289***	***0,7864***	***0,8435***

Table 36: Expert 1 - Résumé des F-mesures sur la base de la classification de 201 nouvelles

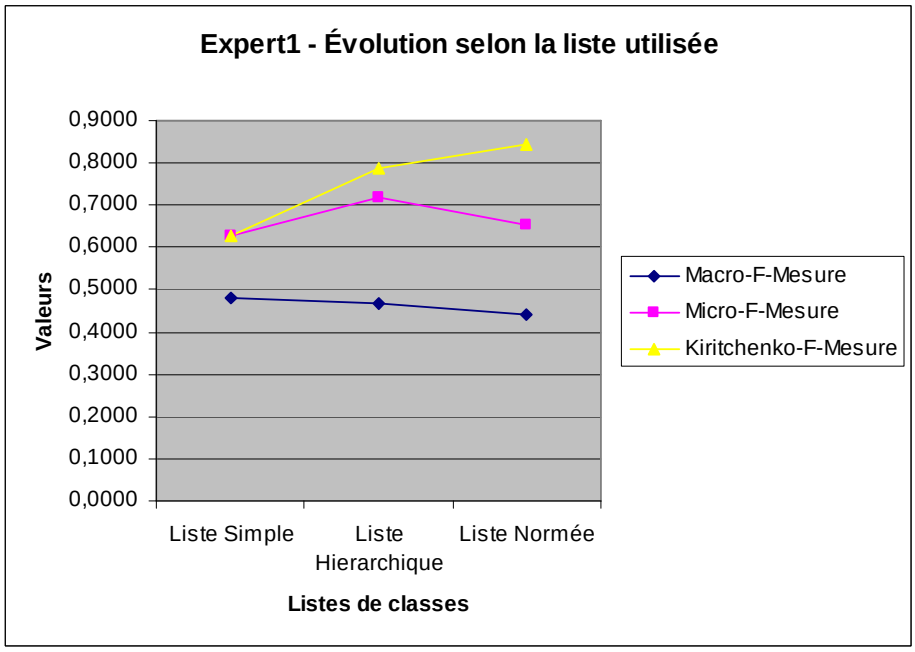


Figure 44: Évolution des résultats de la F-Mesure pour l'expert 1

EXPERT 2:			
Mesures	**Liste Simple**	**Liste Hiérarchique**	**Liste Normée**
Macro-F-Mesure	***0,4532***	***0,4237***	***0,4322***
Micro-F-Mesure	***0,6394***	***0,7211***	***0,6321***
Kiritchenko-F-Mesure	***0,6394***	***0,8063***	***0,8321***

Table 37: Expert 2 - Résumé des F-mesures sur la base de la classification de 201 nouvelles

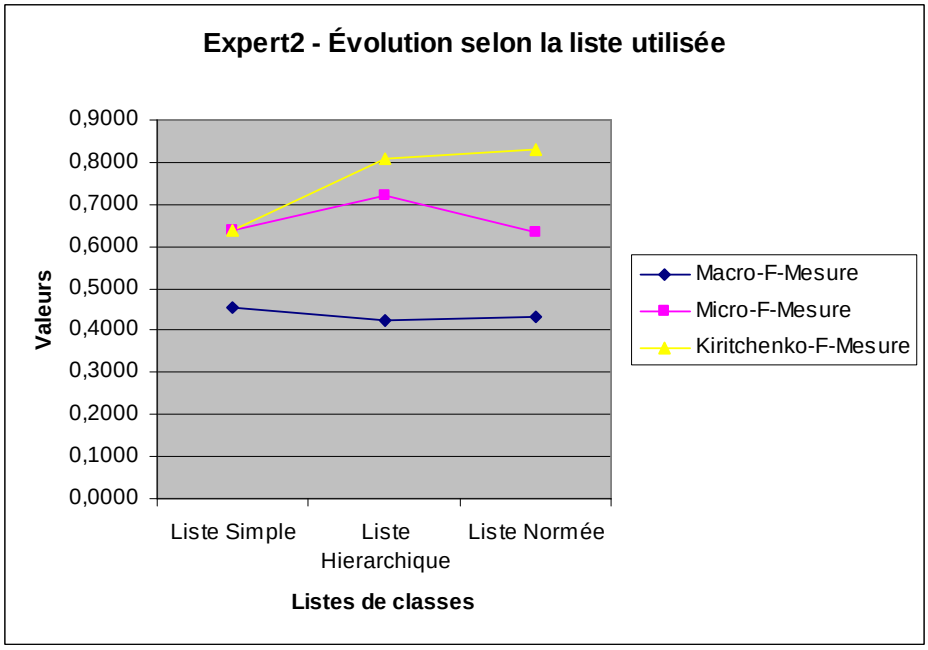


Figure 45: Évolution des résultats de la F-Mesure pour l'expert 2

En analysant les graphes de la Figure 44 et de la Figure 45 nous notons une forme d'amélioration conséquente dans les mesures de Kiritchenko lorsque la liste normée est utilisée. En fait, cela prouve que les erreurs de classification ont un poids plus faible lorsque les classes choisies de façon erronée par ICM sont parentes avec les bonnes classes. Et plus la parenté est proche (même parent) plus l'erreur est moindre. Le contraire est vrai.

K.1.3 Constat global

La classification sur la base de l'entraînement d'ICM par l'expert 1 puis par l'expert 2 a donné des résultats intéressants lorsque la parenté des classes est prise en considération, mais contient néanmoins des baisses de performance lorsqu'il s'agit des mesures autres que celles de Kiritchenko dès qu'on passe d'une liste hiérarchique à une liste normée, quelque soit l'expert. On suppose que le choix des nouvelles a été trop restreint pour pouvoir prouver de façon plus appuyée de l'utilité de la parenté. Pour cette raison, nous avons repris un échantillon plus important sur la base des mêmes conditions (même ensemble d'entraînement et mêmes mesures de performance) afin de vérifier l'amélioration notée ci haut.

K.2 Extension des calculs

Afin d'appuyer le constat précédent montrant l'utilité d'utiliser une forme de parenté dans les classes lors de la classification des nouvelles, nous avons pris un échantillon plus grand constitué des 2 tiers du nombre de nouvelles affectées à des classes dominantes. Nous avons alors refait les calculs pour 402 nouvelles.

K.2.1 Résultats selon les mesures classiques de Sébastiani

EXPERT 1	Liste Simple	Liste Hiérarchique	Liste Normée
Précision	0,5026	0,6247	0,6293
Rappel	0,7876	0,8130	0,6691
F-Mesure	**0,6136**	**0,7065**	**0,6486**

EXPERT 2	Liste Simple	Liste Hiérarchique	Liste Normée
Précision	0,6638	0,7847	0,7515
Rappel	0,6765	0,7074	0,5593
F-Mesure	**0,6701**	**0,7440**	**0,6413**

Table 38: Résultats des mesures de Sébastiani sur la base de 402 nouvelles

Constat : Nous remarquons que la liste normée n'enregistre pas une amélioration dans les performances. Même constat que celui de la section K.1.1.

K.2.2 Résultats selon les mesures élaborées et Kiritchenko

Liste simple			
Mesures globales		**Exp1**	**Exp2**
Macro-moyenne	Macro-précision	0,4157	0,4693
	Macro-Rappel	0,5787	0,4117
	Macro-F-Mesure	**0,4838**	**0,4386**
Micro-moyenne	Micro-précision	0,5026	0,6638
	Micro-Rappel	0,7876	0,6765
	Micro-F-Mesure	**0,6136**	**0,6701**
Kiritchenko	Micro-précision	0,5230	0,6228
	Micro-Rappel	0,7886	0,6569
	Micro-F-Mesure	**0,6289**	**0,6394**

Table 39: Liste simple - Mesures de Kiritchenko suite à la classification de 402 nouvelles

Liste hiérarchique			
Mesures globales		**Exp1**	**Exp2**
Macro-moyenne	Macro-précision	0,4470	0,4698
	Macro-Rappel	0,5207	0,3474
	Macro-F-Mesure	**0,4811**	**0,3995**
Micro-moyenne	Micro-précision	0,6247	0,7847
	Micro-Rappel	0,8130	0,7074
	Micro-F-Mesure	**0,7065**	**0,7440**
Kiritchenko	Micro-précision	0,6956	0,8394
	Micro-Rappel	0,8750	0,8157
	Micro-F-Mesure	**0,7750**	**0,8274**

Table 40: Liste hiérarchique – Mesures de Kiritchenko pour 402 nouvelles

Liste normée			
Mesures globales		**Exp1**	**Exp2**
Macro-moyenne	Macro-précision	0,4874	0,5184
	Macro-Rappel	0,4433	0,3491
	Macro-F-Mesure	**0,4643**	**0,4173**
Micro-moyenne	Micro-précision	0,6293	0,7515
	Micro-Rappel	0,6691	0,5593
	Micro-F-Mesure	**0,6486**	**0,6413**
Kiritchenko	Micro-précision	0,8263	0,9041
	Micro-Rappel	0,8727	0,7696
	Micro-F-Mesure	**0,8489**	**0,8314**

Table 41: Liste normée - Mesures de Kiritchenko suite à la classification de 402 nouvelles

Constat : Selon les explications données à la section K.1.2 nous considérons que les valeurs obtenues sont complexes et puisque la F-Mesure est une base de raisonnement que nous avons adopté dans ce projet, nous avons résumé les mesures en ne gardant que les F-Mesures pour chaque liste et chaque expert (voir la Table 42 et la Table 43). Ces valeurs ont été ensuite utilisées pour tracer un graphe pour chaque expert (voir la Figure 46 et la Figure 47).

E X P E R T 1 :			
Mesures	***Liste Simple***	***Liste Hiérarchique***	***Liste Normée***
Macro-F-Mesure	***0,4838***	***0,4811***	***0,4643***
Micro-F-Mesure	***0,6136***	***0,7065***	***0,6486***
Kiritchenko-F-Mesure	***0,6289***	***0,7750***	***0,8489***

Table 42: Expert 1 - Résumé des F-mesures sur la base de la classification de 402 nouvelles

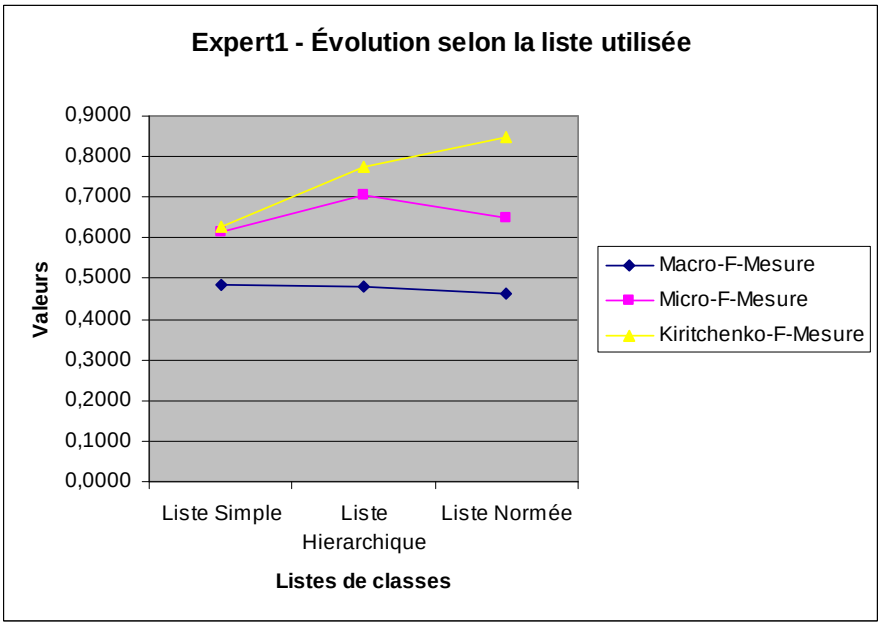


Figure 46: Cas d'un gros échantillon -Évolution des résultats de la F-Mesure pour l'expert 1

E X P E R T 2:

Mesures	*Liste Simple*	*Liste Hiérarchique*	*Liste Normée*
Macro-F-Mesure	***0,4386***	***0,3995***	***0,4173***
Micro-F-Mesure	***0,6701***	***0,7440***	***0,6413***
Kiritchenko-F-Mesure	***0,6394***	***0,8274***	***0,8314***

Table 43: Expert 2 - Résumé des F-mesures sur la base de la classification de 402 nouvelles

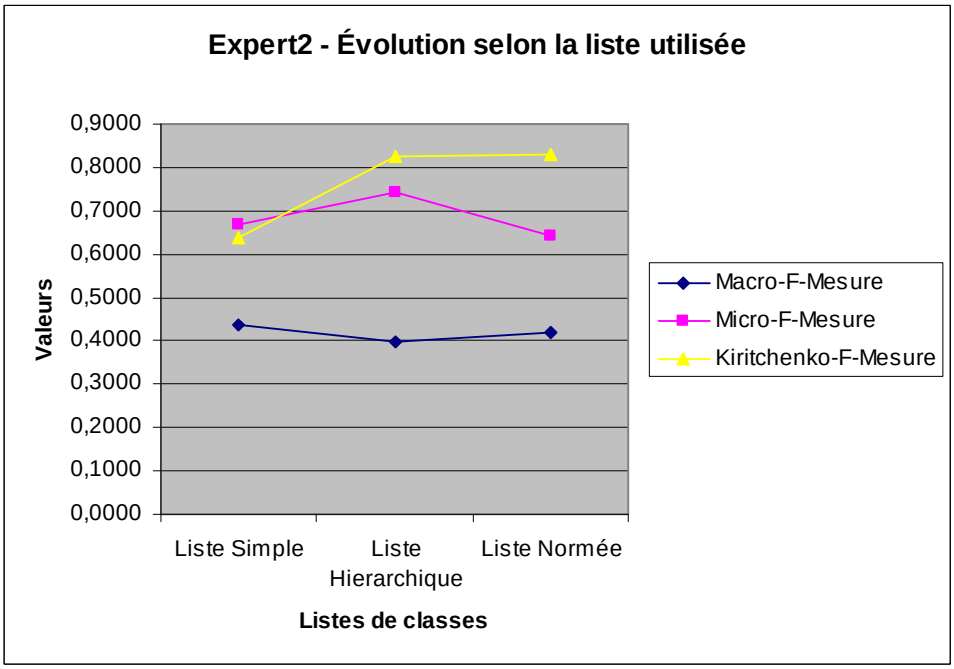


Figure 47: Cas d'un gros échantillon -Évolution des résultats de la F-Mesure pour l'expert 2

En analysant les graphes de la Figure 46 et de la Figure 47 nous notons une forme d'amélioration conséquente dans les mesures de Kiritchenko lorsque la liste normée est utilisée. En fait, cela prouve que les erreurs de classification ont un poids plus faible lorsque les classes choisies de façon erronée par ICM sont parentes avec les bonnes classes. Et plus la parenté est proche (même parent) plus l'erreur est moindre. Le contraire est vrai.

K.2.3 Constat global

Alors que l'échantillon des nouvelles a été étendu dans le but de faire un calcul plus réaliste des performances d'un gros système de classification automatique, nous notons tout de même le même problème qui s'exprime dans le fait que les mesures de classification classiques (autres que Kiritchenko) semblent limitées et incapables de mettre l'accent sur les liens de parenté liant les classes des listes hiérarchique et normée. Le constat est donc le même que pour celui d'un échantillon réduit à la section K.1.3. Pourtant, les mesures de Kiritchenko montraient que ce dernier point influence particulièrement les performances et ne semble pas être en contradiction avec d'autres mesures. En analysant de plus près les listes de classes utilisées et non plus les nouvelles, on a noté un déséquilibre dans leur construction donnant lieu à des erreurs mathématiques et baissant de la sorte les performances du classificateur. Ce déséquilibre étant une piste à étudier, sera expliqué dans la section suivante.

K.3 Correction de la liste normée des classes utilisées

Une fois la classification des 402 nouvelles finalisée sur la base d'une liste normée de 14 classes, nous avons analysé dans le détail le choix des classes fait par ICM et par chaque expert, et nous avons conclu que l'erreur se trouvait dans le fait que 8 des classes feuilles de la liste normée touchaient une partie des nouvelles et non toutes les nouvelles car n'apparaissant pas dans les listes simple et hiérarchique. En fait nous avons comparé en usant d'une probabilité différente qui ne fournissait donc pas le bon résultat. La probabilité concernant le fait qu'une nouvelle quelconque soit affectée à l'une des classes dominantes est de 1/6 (il y a 6 classes feuilles) dans la liste simple et dans la liste hiérarchique. La même nouvelle a une probabilité moins importante face à une liste normée dont les feuilles ne correspondent pas totalement à celles des listes simple et hiérarchique (probabilité de 1/14). D'où un problème de déséquilibre entre la classification des nouvelles sur la base de nos listes simple et hiérarchique (avec 6 classes feuilles) et notre liste normée (avec 14 classes feuilles).

K.3.1 Classification des nouvelles grâce à une liste normée révisée de classes

Afin de corriger l'anomalie des résultats précédents, une nouvelle liste de pré entraînement contenant 203 nouvelles pré classées dans 6 classes feuilles dominantes a été choisie. L'échantillon sélectionné contient 462 nouvelles. La liste normée a été réduite pour ne contenir que les classes apparaissant dans les listes simple et hiérarchique (6 classes) (voir la Figure 48).

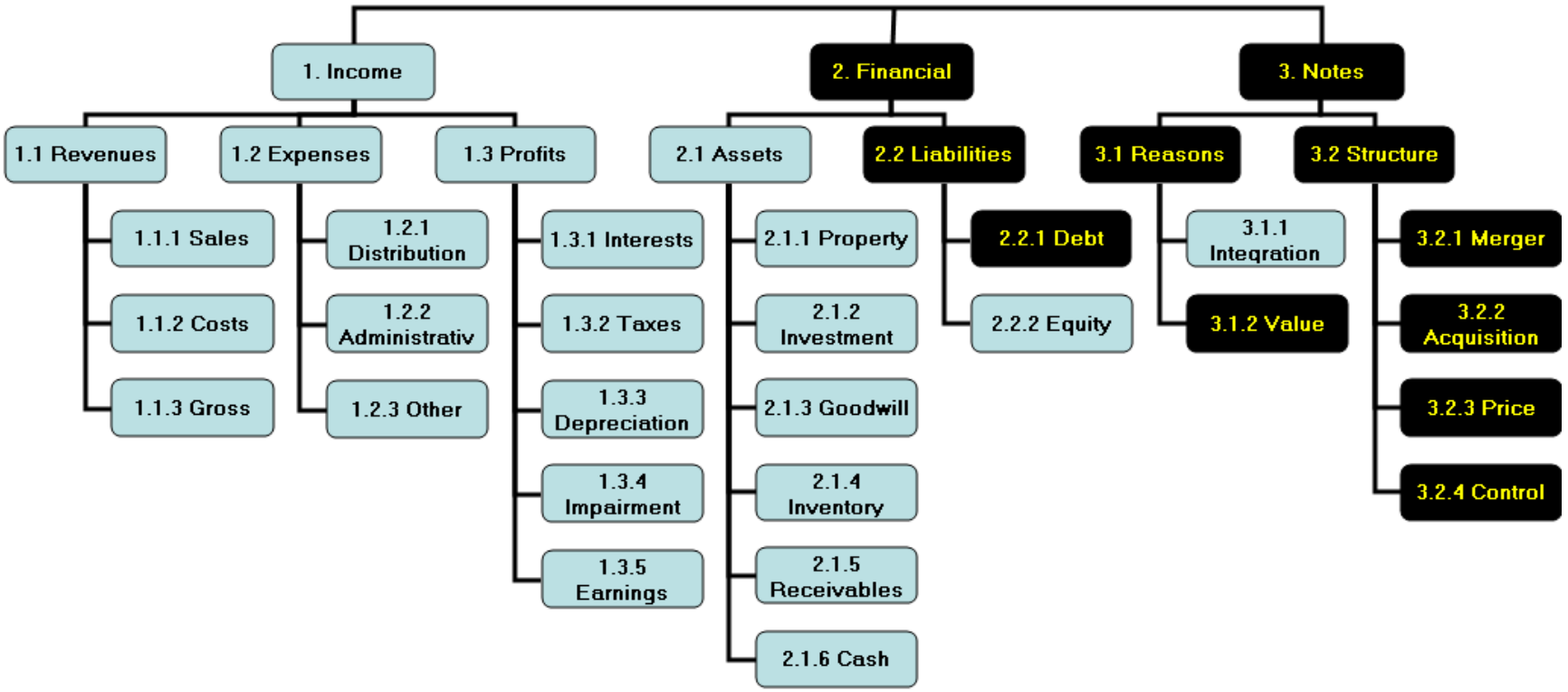


Figure 48: Liste normée réduite (classes sous fond noir)

K.3.2 Résultats selon les mesures classiques de Sébastiani

EXPERT 1	Liste Simple	Liste Hiérarchique	Liste Normée
Précision	0,5870	0,7156	0,8173
Rappel	0,8104	0,8414	0,8407
F-Mesure	**0,6809**	**0,7734**	**0,8288**

EXPERT 2	Liste Simple	Liste Hiérarchique	Liste Normée
Précision	0,6684	0,7351	0,7701
Rappel	0,7474	0,8261	0,8728
F-Mesure	**0,7057**	**0,7779**	**0,8182**

Table 44: Classes révisées - Mesures de Sébastiani suite à la classification de 462 nouvelles

Constat : La mesure considérée étant la F-Mesure qui permet d'équilibrer le poids du rappel et de la précision, nous avons constaté ce qui suit pour chaque expert (afin de mieux comprendre le raisonnement ayant permit de déduire la F-Mesure, se référer à l'annexe C, à la section C.2):

3. La valeur enregistrée par la liste hiérarchique est plus importante que celle de la liste simple (exemple : 0,7734 par rapport à 0,6809 pour l'expert 1)
4. La valeur enregistrée par la liste normée est plus importante que celle de la liste hiérarchique (exemple : 0,8288 par rapport à 0,7734 pour l'expert 1)

Le fait de ne garder dans la liste normée que les classes feuilles similaires à celles des 2 autres classes, a permit d'améliorer de façon évidente les résultats de la classification des nouvelles sur la base d'une liste normée. En fait, on constate que l'amélioration se poursuit et devient plus importante lorsque les niveaux d'hiérarchie sont plus importants (1 seul niveau pour la liste simple, 2 niveaux pour la liste hiérarchique et 3 niveaux pour la liste normée). La parenté entre les classes ainsi que l'utilisation d'une ontologie spécifique au domaine ont permit d'améliorer les résultats d'une classification hiérarchique de nouvelles dans le cas de mesures classiques de Sébastiani.

K.3.3 Résultats selon les mesures élaborées et Kiritchenko

Liste simple			
Mesures globales		**Exp1**	**Exp2**
Macro-moyenne	Macro-précision	0,4801	0,4369
	Macro-Rappel	0,5589	0,4467
	Macro-F-Mesure	**0,5165**	**0,4417**
Micro-moyenne	Micro-précision	0,5870	0,6684
	Micro-Rappel	0,8104	0,7474
	Micro-F-Mesure	**0,6809**	**0,7057**
Kiritchenko	Micro-précision	0,5870	0,6684
	Micro-Rappel	0,8104	0,7474
	Micro-F-Mesure	**0,6809**	**0,7057**

Table 45: Liste simple de classes révisées - Mesures de Kiritchenko pour 462 nouvelles

Liste hiérarchique			
Mesures globales		**Exp1**	**Exp2**
Macro-moyenne	Macro-précision	0,5030	0,4968
	Macro-Rappel	0,5083	0,5365
	Macro-F-Mesure	**0,5056**	**0,5159**
Micro-moyenne	Micro-précision	0,7156	0,7351
	Micro-Rappel	0,8414	0,8261
	Micro-F-Mesure	**0,7734**	**0,7779**
Kiritchenko	Micro-précision	0,7900	0,8368
	Micro-Rappel	0,8961	0,8830
	Micro-F-Mesure	**0,8397**	**0,8593**

Table 46: Liste hiérarchique de classes révisées -Mesures de Kiritchenko pour 462 nouvelles

Liste normée			
Mesures globales		**Exp1**	**Exp2**
Macro-moyenne	Macro-précision	0,5308	0,5592
	Macro-Rappel	0,4637	0,5737
	Macro-F-Mesure	**0,4950**	**0,5664**
Micro-moyenne	Micro-précision	0,8173	0,7701
	Micro-Rappel	0,8407	0,8728
	Micro-F-Mesure	**0,8288**	**0,8182**
Kiritchenko	Micro-précision	0,7787	0,8024
	Micro-Rappel	0,7869	0,9084
	Micro-F-Mesure	**0,7828**	**0,8521**

Table 47: Liste normée de classes révisées - Mesures de Kiritchenko pour 462 nouvelles

Constat : Selon les explications données à la section K.1.2 nous considérons que les valeurs obtenues sont complexes et puisque la F-Mesure est une base de raisonnement que nous avons adopté dans ce projet, nous avons résumé les mesures en ne gardant que les F-Mesures pour chaque liste et chaque expert (voir la Table 48 et la Table 49). Ces valeurs ont été ensuite utilisées pour tracer un graphe pour chaque expert (voir la Figure 49 et la Figure 50).

E X P E R T 1 :			
Mesures	***Liste Simple***	***Liste Hiérarchique***	***Liste Normée***
Macro-F-Mesure	***0,5165***	***0,5056***	***0,4950***
Micro-F-Mesure	***0,6809***	***0,7734***	***0,8288***
Kiritchenko-F-Mesure	***0,6809***	***0,8397***	***0,7828***

Table 48: Expert 1 - F-mesures pour 462 nouvelles avec une liste de classes révisées

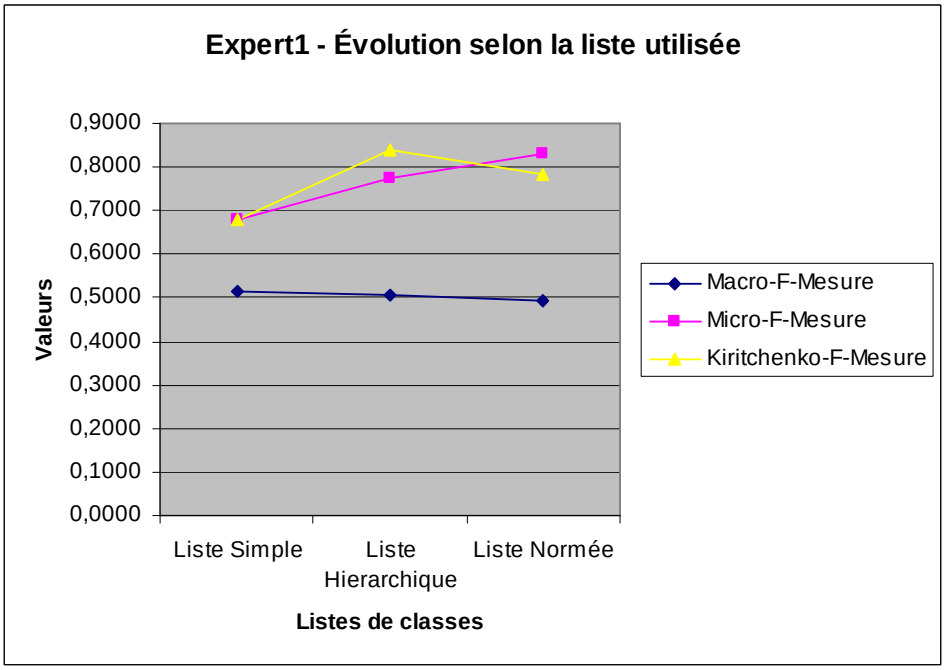


Figure 49: Évolution de la F-Mesure avec une liste normée à 6 classes pour l'expert 1

E X P E R T 2 :

Mesures	*Liste Simple*	*Liste Hiérarchique*	*Liste Normée*
Macro-F-Mesure	***0,4417***	***0,5159***	***0,5664***
Micro-F-Mesure	***0,7057***	***0,7779***	***0,8182***
Kiritchenko-F-Mesure	***0,7057***	***0,8593***	***0,8521***

Table 49: Expert 2 - F-mesures pour 462 nouvelles avec une liste de classes révisées

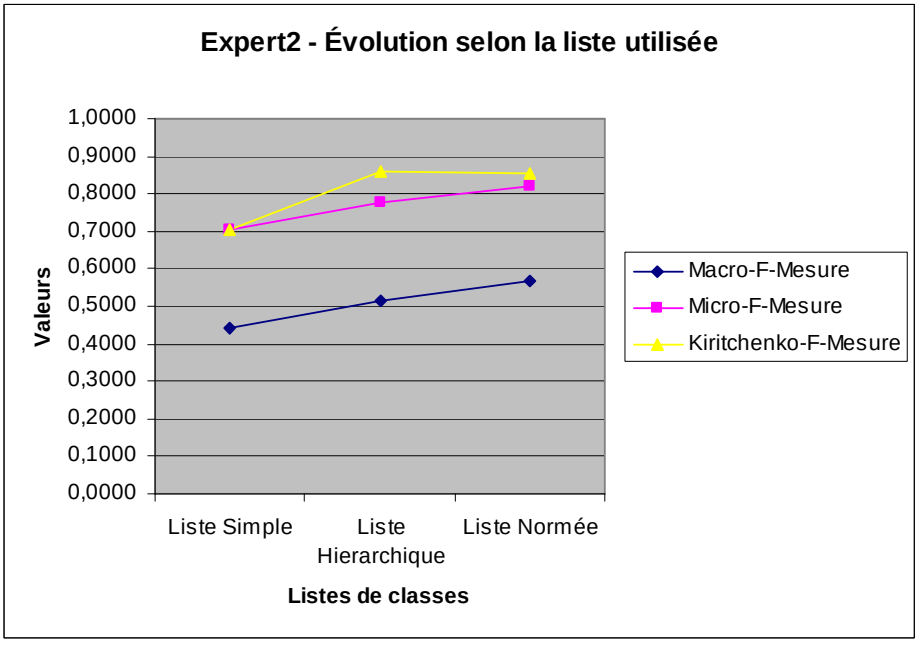


Figure 50: Évolution de la F-Mesure avec une liste normée à 6 classes pour l'expert 2

En analysant les graphes de la Figure 49 et de la Figure 50 nous notons une forme d'amélioration conséquente dans les mesures de micro et macro moyennes lorsque les niveaux d'hiérarchie de la liste de classes utilisée augmente (de la liste simple à la hiérarchique, et de la liste hiérarchique à la liste normée). Mais ce qui n'était pas prévu était la diminution des performances de la classification lorsque la liste normée est utilisée dans les mesures de Kiritchenko.

K.3.4 Constat global

Dans ce cas de figure, la parenté a pourtant prouvé dans le cas des mesures de Sébastiani qu'il y avait une amélioration significative dans les résultats de la performance alors que les mesures hiérarchiques de Kiritchenko ont noté une baisse de performance dans le cas d'une liste normée. Ce cas de figure étant complexe à développer, il est nécessaire de répondre à la question en réduisant le problème à un ensemble de cas plus simples à traiter. Cette façon de faire nous permettra d'analyser nouvelle par nouvelle les anomalies de la chute de performance d'une liste normée dans le cas de mesures de Kiritchenko spécifiques aux classes parentes. La section K.4 nous fournira en détail la méthode adoptée et les résultats obtenus. Cette méthode va se concentrer sur le raisonnement du classificateur plutôt que sur des calculs qui pourraient mettre de côté la valeur d'une classification améliorée et/ou enrichie pour une liste normée par rapport aux listes simple et hiérarchique.

K.4 Analyse du raisonnement logique du classificateur

Afin de mieux comprendre la façon dont ICM a classifié les nouvelles sur la base d'une liste normée et révisée de classes, nous avons utilisé la méthode enrichie des 4 cas de figure présentée à la section 3.7. En classifiant les 462 nouvelles précédemment sélectionnées sur la base des listes simple, hiérarchique et normée, nous avons noté qu'ICM avait amélioré la classification de certaines nouvelles mais a mal classifié d'autres. La méthode des 4 cas de figure a permit de montrer de façon plus détaillée ces différents cas de classification en dressant des ensembles de nouvelles dont le lien commun concerne la façon dont ces dernières ont été classifiées. Ainsi, des ensembles de nouvelles bien, mal et mieux classifiées sont présentés dans la Table 50 pour l'expert 1 et la Table 51 pour l'expert 2.

Statistiques du nombre de nouvelles mieux, bien, moins bien et mal classées par rapport à l'expert 1

Signification	Code	Totaux		
		LS-LH	LS-LN	LH-LN
Amélioration	AT	3	1	1
Amélioration	AP	1	2	1
Amélioration	NBCPE	15	46	37
Bonne classification stable	NBC	99	89	98
Enrichissement	ACE	4	1	2
Enrichissement	ES	212	164	163
Enrichissement	EA	0	0	0
Enrichissement	ED	0	0	0
Stabilité négative	NMC	24	24	23
Stabilité négative	NPBC	96	96	103
Diminution légère de la performance	DP	8	35	29
Diminution grave de la performance	DT	0	4	5

Résumé			
	LS-LH	LS-LN	LH-LN
Stabilité	219	209	224
Amélioration	235	214	204
Diminution	8	39	34
fraction	0,97	0,85	0,86

Résumé	Totaux		
	LS-LH	LS-LN	LH-LN
Amélioration et bonne classification stable	118	138	137
Enrichissement	216	165	165
Stabilité négative	120	120	126
Diminution de la performance	8	39	34

Table 50: Classification d'un échantillon de 462 nouvelles par rapport à l'expert 1

Statistiques du nombre de nouvelles mieux, bien, moins bien et mal classées par rapport à l'expert 2

Signification	Code	Totaux		
		LS-LH	LS-LN	LH-LN
Amélioration	AT	14	25	20
Amélioration	AP	106	101	3
Amélioration	NBCPE	10	7	5
Bonne classification stable	NBC	158	132	136
Enrichissement	ACE	25	52	36
Enrichissement	ES	102	106	130
Enrichissement	EA	2	2	3
Enrichissement	ED	8	11	5
Stabilité négative	NMC	20	16	17
Stabilité négative	NPBC	7	6	103
Diminution légère de la performance	DP	4	3	4
Diminution grave de la performance	DT	6	1	0

Résumé	LS-LH	LS-LN	LH-LN
Stabilité	185	154	256
Amélioration	267	304	202
Diminution	10	4	4
fraction	0,96	0,99	0,98

Résumé	Totaux		
	LS-LH	LS-LN	LH-LN
Amélioration et bonne classification stable	288	265	164
Enrichissement	137	171	174
Stabilité négative	27	22	120
Diminution de la performance	10	4	4

Table 51: Classification d'un échantillon de 462 nouvelles par rapport à l'expert 2

Constat global

Si nous analysons les tables ci-haut, nous constatons qu'à la ligne fraction les résultats montrent le niveau de stabilité et d'amélioration ainsi que de diminution.

Si nous considérons que la stabilité ne peut influencer les performances d'un système de classification et que l'amélioration prouve son efficacité, nous avons calculé la fraction qui va permettre de comprendre le pourcentage des classifications améliorées par rapport au reste des classifications.

La fraction est représentée par la formule suivante :

$$Fraction = (Liste\ des\ classifications\ améliorées) / (Liste\ des\ classifications\ améliorées + Liste\ des\ classifications\ diminuées) \quad (18)$$

Cependant, les résultats montrent que le taux de diminution des performances modifie les résultats globaux en baissant le niveau de performance général. Ce constat s'applique surtout à l'expert 1 lorsque le passage vers une liste normée est effectué causant une forme biaisée des résultats.
En analysant mieux les classes choisies par l'expert 1, dans le cas des nouvelles en diminution de performance, une possibilité de mauvaise interprétation des nouvelles est apparue. Afin d'étudier cette piste de plus près, nous avons choisi de récolter toutes les nouvelles en diminution de performance, pour l'expert 1 et pour l'expert 2, et de les ré classifier manuellement par un expert professionnel du domaine.
Les résultats de cette classification manuelle experte seront, ensuite, injectés dans les classifications précédentes, en remplacement de celles biaisées, puis les mesures recalculées.

K.5 Correction de la classification des nouvelles en diminution de performance

En analysant les résultats de la classification des 462 nouvelles précédentes contenant 55 nouvelles ré classifiées manuellement par notre expert du domaine dont 42 nouvelles en diminution de performance issues de la classification de l'expert 1 et 13 de celle de l'expert 2, nous avons enregistré des résultats satisfaisants que nous pouvons consulter à travers la Table 52.

Constat global

Si nous analysons les tables ci bas, nous constatons qu'à la ligne fraction les résultats montrent le niveau de stabilité et d'amélioration ainsi que de diminution. Cette dernière est beaucoup moins importante par rapport aux classifications précédentes (spécialement pour l'expert 1).
Cela prouve qu'ICM a pris de meilleures décisions en se basant sur un entraînement normé issu de l'accord des experts du domaine.
En fait, cela prouve que la diminution n'était pas causée par la liste normée utilisée mais plutôt par une mauvaise compréhension humaine. Cette marge d'erreur est une chose courante dans le domaine de la classification. Pourtant, le fait d'entraîner le classificateur commercial a permis de détecter cette anomalie et nous a permis de prouver de façon satisfaisante que l'amélioration d'une classification quelconque est surtout basée sur un bon entraînement et une liste de classes étudiée et organisée de façon ontologique.

La marge d'erreur est liée à la façon dont le raisonnement du classificateur a été orienté grâce à son entraînement.
Ce qui nous permet de prouver que les classifications sur la base d'une liste normée par rapport aux autres listes enregistre des performances timides mais néanmoins conséquentes.

Expert 1

Avant

Résumé			
	LS-LH	LS-LN	LH-LN
Stabilité	219	209	224
Amélioration	235	214	204
Diminution	8	39	34
fraction	0,97	0,85	0,86

Après

Résumé			
	LS-LH	LS-LN	LH-LN
Stabilité	221	219	231
Amélioration	235	230	222
Diminution	6	13	9
fraction	0,98	0,95	0,96

Expert 2

Avant

Résumé			
	LS-LH	LS-LN	LH-LN
Stabilité	185	154	256
Amélioration	267	304	202
Diminution	10	4	4
fraction	0,96	0,99	0,98

Après

Résumé			
	LS-LH	LS-LN	LH-LN
Stabilité	186	159	257
Amélioration	267	299	202
Diminution	9	4	3
fraction	0,97	0,99	0,99

Table 52: Reclassification experte manuelle de 55 nouvelles en diminution de performance

BIBLIOGRAPHIE

1. Raghavan, H., *Tandem learning: A learning framework for document categorization*, in *Computer Science*. 2007, University of Massachusetts Amherst: Massachusetts. p. 173.
2. Feldman, R., *Mining the Biomedical Literature using Semantic Analysis and Natural Language Processing Techniques, a Link Analysis Approach.* International conference on Intelligent Systems for Molecular Biology (ISMB). 2003.
3. FÜRST, F., *L'ingénierie ontologique*, in *Ingénierie des Connaissances*. 2002, IRIN, Université de Nantes: Nantes, France. p. 38.
4. Teller, P., *Formalisation des normes comptables : vers une ontologie des notions de comptabilité*, ed. H. INFORSID, Tunisie. 2006.
5. DUFRENE, C.L.P., *XBRL - la solution pour une ontologie des informations financières.* OTC-Conseil, 2009 (lettre num 38 avril 2009).
6. IASB. *International Financial Reporting Standards - The IFRS XBRL Taxonomy Illustrated*. 2009 [cited feb 2009]; Available from: http://eifrs.iasb.org/eifrs/Taxonomy?type=r&lang=en.
7. Bazire, S. and M.-N. Maffon, *Impacts de la mise en place des normes IFRS sur les capitaux propres*, in *CNAM Paris - Promotion 2005*. p. 119.
8. Deloitte. *Guide to IFRS 3 and IAS 27 for Business Combinations*. 2008; Available from: http://www.iasplus.com/dttpubs/0807ifrs3guide.pdf.
9. KPMG. *IFRS 3 Business Combinations*. 2004 [cited feb 2009]; Available from: http://www.kpmg.fi/Binary.aspx?Item=1399.
10. NIST. *The Text Retrieval Conference (TREC)*. 2008 [cited 15th october 2008]; Available from: http://trec.nist.gov/overview.html.
11. Ceci, M. and D. Malerba, *Classifying web documents in a hierarchy of categories: a comprehensive study*. Journal of Intelligent Information Systems, 2007. (Vol 28, Issue 1): p. 37-78.
12. Kiritchenko, S., et al., *Learning and Evaluation in the Presence of Class Hierarchies: Application to Text Categorization*, in *Lecture Notes in Computer Science - LNCS - Advances in Artificial Intelligence*. 2006, Springer: Berlin. p. 395-406.
13. Li, T., S. Zhu, and M. Ogihara, *Hierarchical document classification using automatically generated hierarchy*. Journal of Intelligent Information Systems, 2007. (Vol 29, Issue 2): p. 211-230.
14. Conway, M., et al., *Classifying disease outbreak reports using n-grams and semantic features*. International Journal of Medical Informatics, 2009. (In press).
15. Wang, T. and B.C. Desai. *Document Classification with ACM Subject Hierarchy*. in *Canadian Conference on Electrical and Computer Engineering (CCECE)* 2007. Vancouver.
16. He, J., et al., *Categorizing software engineering knowledge using a combination of SWEBOK and text categorization*, in *Lecture Notes in Computer Science - LNCS - Advances in Artificial Intelligence*. 2007, Springer: Berlin. p. 675-681.
17. Sokolova, M. and G. Lapalme, *A systematic analysis of performance measures for classification tasks*. Information Processing and Management, 2009 (Vol 45, Issue 4): p. 427-437.

18. Lewis, D.D., et al. *RCV1-v2/LYRL2004: The LYRL2004 Distribution of the RCV1-v2 Text Categorization Test Collection (12-Apr-2004 Version).* 2004b [cited 2009]; Available from: http://jmlr.csail.mit.edu/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm.
19. Lewis, D.D., et al., *RCV1: A new benchmark collection for text categorization research.* Journal of Machine Learning Research, 2004a. (Vol 5, Issue Dec): p. 361 - 397.
20. Haleblian, J., et al., *Taking stock of what we know about mergers and acquisitions: A review and research agenda.* Journal of Management, 2009. (Vol 35, Issue 3): p. 469-502.
21. Rhéaume, L. and H.S. Bhabra, *Value creation in information-based industries through convergence: A study of U.S. mergers and acquisitions between 1993 and 2005.* Information & Management, 2008. (Vol 45, Issue 5): p. 304-311.
22. Wang, L. and E.J. Zajac, *Alliance or acquisition? a dyadic perspective on interfirm resource combinations.* Strategic Management Journal, 2007. (Vol 28, Issue 13): p. 1291-1317.
23. Nam, C., et al., *Stock market reaction to mergers and acquisitions in anticipation of a subsequent related significant event: Evidence from the Korean telecommunications industry.* Review of Pacific Basin Financial Markets and Policies, 2005. (Vol 8, Issue 2): p. 185-200.
24. Swaminathan, V., F. Murshed, and J. Hulland, *Value creation following merger and acquisition announcements: The role of strategic emphasis alignment.* Journal of Marketing Research, 2008. (Vol 45, Issue 1): p. 33-47.
25. Wan, W.P. and D.W. Yiu, *From crisis to opportunity: environmental jolt, corporate acquisitions, and firm performance.* Strategic Management Journal, 2009. (Vol 30, Issue 7): p. 791-801.
26. Balaban, E. and C.T. Constantinou, *Volatility clustering and event-induced volatility: Evidence from UK mergers and acquisitions.* European Journal of Finance, 2006. (Vol 12, Issue 5): p. 449-453.
27. Cloodt, M., J. Hagedoorn, and H. Van Kranenburg, *Mergers and acquisitions: Their effect on the innovative performance of companies in high-tech industries.* Research Policy, 2006. (Vol 35, Issue 5): p. 642-654.
28. Homburg, C. and M. Bucerius, *Is speed of integration really a success factor of mergers and acquisitions? An analysis of the role of internal and external relatedness.* Strategic Management Journal, 2006. (Vol 27, Issue 4): p. 347-367.
29. Sébastiani, F., *Machine Learning in Automated Text Categorization.* ACM Computing Surveys, 2002. (Vol 34, Issue 1): p. 1-47.
30. Rijsbergen, v., *Information retrieval / c.j.* 1979: London ; Toronto : Butterworths, 1979.
31. Aseervatham, S., *Apprentissage à base de Noyaux Sémantiques pour le Traitement de Données Textuelles.* 2008. p. 220.
32. Fürst, F., *"L'ingénierie ontologique"*, in *"Ingénierie des Connaissances"*. 2002, IRIN, Université de Nantes: Nantes, France. p. 38.
33. Giuseppe, B.A. and P. Maccarrone, *IFRSs and accounting for intangible assets: The Telecom Italia case.* Journal of Intellectual Capital, 2007. (Vol 8, Issue 2): p. 306-328.
34. Kimbrough, M.D., *The influences of financial statement recognition and analyst coverage on the market's valuation of R&D capital.* Accounting Review, 2007. (Vol 82, Issue 5): p. 1195-1225.

35. Martinez-Jerez, F.A., *Governance and merger accounting: Evidence from stock price reactions to purchase versus pooling.* European Accounting Review, 2008. (Vol 17, Issue 1): p. 5-35.
36. IBM, *Classification Module: Classification Workbench User's Guide, V8.6 (SC18-9878-02)*. 2007, IBM Publication Center.
37. Grosse, M. *Classify your enterprise content and manage taxonomies using IBM Classification Module*. 2007 [cited 1st November 2008]; Available from: http://www.ibm.com/developerworks/db2/library/techarticle/dm-0712grosse/.
38. IBM. *IBM Classification Module Version 8.5 information roadmap*. 2007 [cited 15th june 2008]; Available from: http://www-01.ibm.com/support/docview.wss?rs=3376&uid=swg27010905.
39. Moskovitch, R., et al., *Multiple hierarchical classification of free-text clinical guidelines.* Artificial Intelligence in Medicine, 2006. (Vol 37, Issue 3): p. 177-190.
40. Koller, D. and M. Sahami, *Hierarchically classifying documents using very few words*. 1997, Stanford InfoLab.
41. McCallum, A., et al., *Improving Text Classification by Shrinkage in a Hierarchy of Classes*, in *Proceedings of the Fifteenth International Conference on Machine Learning*. 1998, Morgan Kaufmann Publishers Inc.
42. Piwowarski, B., L. Denoyer, and P. Gallinari, *Un modèle pour la recherche d'information sur des documents structurés*, S.-M. 6èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT 2002), France, Editor. 2002.
43. Kiritchenko, S., *Hierarchical text categorization and its application to bioinformatics*. 2006, University of Ottawa. p. 187.
44. Kiritchenko, S., S. Matwin, and A. Famili, *hierarchical text categorization as a tool of associating genes with gene ontology codes.* 2004. (Proceedings of the 2nd European Workshop on Data Mining and Text Mining for Bioinformatics): p. 26–30.
45. Kunchukuttan, A., *Evaluation of Information Retrieval Systems*, in *Department of Computer Science and Engineering*. 2006, Indian Institute of Technology, Bombay: Mumbai. p. 27.
46. Lespinasse, K., *TREC Une conférence pour l'évaluation des systèmes de recherche d'information.* Sciences de l'information, 1997: p. 77-81.
47. Yang, Y., *An Evaluation of Statistical Approaches to Text Categorization.* Information Retrieval, 1999. (Vol 1, Issue 1): p. 69-90.
48. Mille, A. *Ontologies.* [cited March 2009]; Available from: http://liris.cnrs.fr/alain.mille/enseignements/DEA-ECD/ontologies/.
49. Guarino, N., *Formal Ontology in Information Systems: Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy*. 1998: IOS Press. 337.
50. Gruber, T.R., *A translation approach to portable ontology specifications.* Knowl. Acquis., 1993. **5**(2): p. 199-220.
51. Psyché, V., O. Mendes, and J. Bourdeau, *Apport de l'ingénierie ontologique aux environnements de formation à distance*. 2003.
52. Stumme, G., et al., *The {Karlsruhe} View on Ontologies*. 2003.
53. Hubbard, S. *XBRL s'en vient!* septembre 2003 [cited 1st nov 2008]; Available from: http://www.camagazine.com/index.cfm/ci_id/16433/la_id/2.htm.
54. Jarry, E., *XBRL: Introduction à la technologie XML des rapports financiers*. Livre blanc - Révision 1.0 - Janvier 2006 ed. Livre blanc - Introduction à XBRL, ed. X. France. 2006, Paris: XBRL France.

55. Canada, X. *The CSA XBRL Voluntary Filing Program* 2008 [cited 1st dec 2008]; Available from: http://www.xbrl.ca/e/CSAVP.html.
56. Canada, X. *A small example of XBRL* 2008 [cited 1st dec 2008]; Available from: http://xbrl.org/nmpxbrl.aspx?id=44.