Chapitre V. Classification des textes prophétiques basée sur l'algorithme des arbres de décision*

1. Introduction

Ce chapitre présente les résultats d'évaluation de l'algorithme des arbres de décision dans la tâche de classification des textes arabes. Les Expérimentations sont dérouleés sur deux corpus en langue arabe. Les résultats montrent que l'utilisation hybride du seuil de la fréquence des termes *TFT (Term Frequency Thresholding)* avec le critère du gain de l'information *IG (Information Gain),* déjà existant dans l'algorithme des arbres de décision lui-même, forme l'approche préférable pour la sélection des attributs. L'étude montre que l'amélioration de notre classificateur basé sur les arbres de décision donne de très bons résultats du point de vue efficacité et performance. L'exactitude générale est autours du 93% pour le corpus scientifique et de 91% pour le corpus littéraire. Nous démontrons aussi que l'augmentation de la taille de l'ensemble d'apprentissage et la nature du corpus ont une grande influence sur l'amélioration des performances du classificateur. Ainsi la réduction des erreurs de classification nous oblige de choisir soigneusement les documents de notre corpus qui doivent appartenir à des catégories les plus thématiquement divergentes.

2. Classification par les arbres de décision

Les arbres de décision peuvent être classés parmi les techniques de l'apprentissage supervisé. Ils sont populaires et pratiques. Les deux algorithmes les plus connus et utilisés sont CART (Classification And Regression Trees [Breiman et. al, 1984]) et C5 (version la plus récente après ID3 et C4.5 [Quinlan, 1993]). Ces algorithmes sont performants et génèrent des procédures de classification exprimables sous forme de règles. Pour utiliser un arbre de décision, c'est-à-dire classer un enregistrement, il suffit de descendre dans l'arbre selon les réponses aux différents tests pour l'enregistrement considéré. Les règles des systèmes construits sont exhaustives et mutuellement exclusives. Cela signifie que pour tout enregistrement une et une seule règle s'applique.

2.1. Algorithme arbre de Décisions

Le schéma général de l'algorithme *arbre de décision* est donné ci-dessous, L'idée centrale est de diviser récursivement et le plus efficacement possible les exemples de l'ensemble d'apprentissage par des tests définis à l'aide des attributs jusqu'à ce que l'on obtienne des sous ensembles d'exemples ne contenant (presque) que des exemples appartenant tous à une même classe. Dans toutes les méthodes, on trouve les trois opérations **(1, 2, 3)** cités ci-dessous. Les algorithmes des arbres de décision diffèrent par les choix effectués par ces opérations, c'est-à-dire sur le choix d'un test (par exemple, utilisation du gain et de la fonction entropie) et le critère d'arrêt (soit quand il faut arrêter la croissance de l'arbre, soit quand décider si un nœud est terminal).

* Une grande partie de ce chapitre a été publiée dans les articles suivants :

1. F.Harrag, A. Hamdi-Cherif, K. Dellidj, et Y. Benyahia, Arabic Text Categorization Based On the Decision Trees Algorithm, *Journal of Computer Science and Engineering in Arabic*, ISSN 1936-0525, Vol.1, N.2, pp. 72-96, 2007.
2. F. Harrag, E. El-Qawameh and A. Hamdi-Cherif, Performance of Decision Trees on Arabic Text Categorization, *Journal of Digital Information Management*, ISSN 0972-7272, Vol.7, N.6, pp. 377-382, 2009.

Donnée : un échantillon *S* de *m* enregistrements classés *(x, c(x))*
Initialisation : A :=arbre vide ;
Noeud_courant := racine ;
Échantillon courant := S
Répéter **1-** Décider **Si** (Noeud_courant est terminal) **Alors**
2- Étiqueter le noeud courant par une feuille.
Sinon **3-** Sélectionner un test ; Créer les fils.
Définir les échantillons sortants du nœud.
Fin de si
Noeud_courant := un noeud non encore étudié de A.
Échantillon_courant : échantillon atteignant noeud_courant.
Jusqu'à obtenir l'arbre de décision.

Algorithme 5.1. Algorithme de classification par *arbre de décision*

2.2. Calcul de L'Entropie

L'entropie est une valeur numérique quantifiant la notion de dispersion d'information (accroissement du désordre). Soit S un système partitionné en k sous systèmes alors son entropie E(S) est:

$$Entropie\ (S) = \sum_{k=1} p\left(\frac{k}{p}\right) \cdot \log\left(p\left(\frac{k}{p}\right)\right) \qquad (5.1)$$

Où :

P (k/p)=N (k/p)/N (p) : La proportion d'éléments de classe k à la position *p*.

N (p) : Est le cardinal de l'ensemble des exemples associés au nœud *p*.

N (k /p) : Est le cardinal de l'ensemble des exemples associés à *p* qui sont de classe *k*.

2.3. Le Choix de l'Attribut Gagnant

Le gain est l'entropie du nœud actuel moins la somme pondéré de l'entropie des nœuds crées.

$$Gain(p,t) = i(p) - \sum_{j=1}^{n} p_j \cdot i(p_j) \qquad (5.2)$$

i(p):Représente l'entropie actuelle du noeud *p*.

On souhaite obtenir l'entropie la plus faible car cela signifie qu'un grand nombre d'éléments appartiennent à la classe, on cherche donc à obtenir le gain maximum.

2.4. Un Exemple pour démonstration

Dans cet exemple démonstratif, nous exposons la démarche de construction d'un arbre de décision. La figure 5.1 montre un exemple de l'arbre de décision résultant de l'application de l'algorithme précédent sur un ensemble de documents arabes. Le premier niveau de l'arbre "la racine" contient l'ensemble initial de tous les termes extraits à partir des textes de ces documents. Le dernier niveau "les feuilles" représente l'ensemble des classes dans les quelles sont affectés ces documents. Le niveau intermédiaire représente l'ensemble des nœuds résultants de l'opération de division basée sur le calcul de l'entropie et le gain de l'information pour tous les termes de la racine de l'arbre. Le mot "Oui" est utilisé pour représenter l'apparition du terme *T* dans le document *D* et le mot "Non" pour représenter l'absence du terme *T* dans le document *D.*

Soit *S* un échantillon de 50 documents divisé sur 2 classes, 38 documents pour la classe "*Historique*" et 12 documents pour la classe "*Sciences islamiques*". L'entropie du nœud initial est calculée par la formule suivante :

$$Entropie(s) = -\left(\frac{38}{50}\right) \times \log\left(\frac{38}{50}\right) - \left(\frac{12}{50}\right) \times \log\left(\frac{12}{50}\right) \quad (5.3)$$

Ensuite on calcule le gain d'information pour l'ensemble des tous les termes ou "Attributs", on a 14 cas d'apparition pour l'attribut "*Fataha*, فتح ", donc le gain d'information pour cet attribut est calculé selon la formule :

$$Gain(S,' Fataha') = i(S) - \left(\frac{36}{50}\right) \times i(S_{SI}) - \left(\frac{14}{50}\right) \times i(S_H) = 0.189 \quad (5.4)$$

Où : ***SI*** *: Sciences islamiques et* ***H*** *: Historique.*

Après calcul, le gain d'information pour l'attribut "*Arakha*, أرخ " est égal à 0.1339, et pour l'attribut "*Kharâa*, خرع " est égal à 0.0567, et pour l'attribut "*Rassala*, رسل" est égal à 0.0567. On trouve que l'attribut "*Fataha*, فتح " a eu la plus grande valeur du gain qui est égale à 0.189 donc il devient la racine de l'arbre. De la même manière on continue la construction de l'arbre de décision en se basant sur le calcul du gain d'information pour tous les attributs et on obtient l'arbre de décision de la figure 5.1.

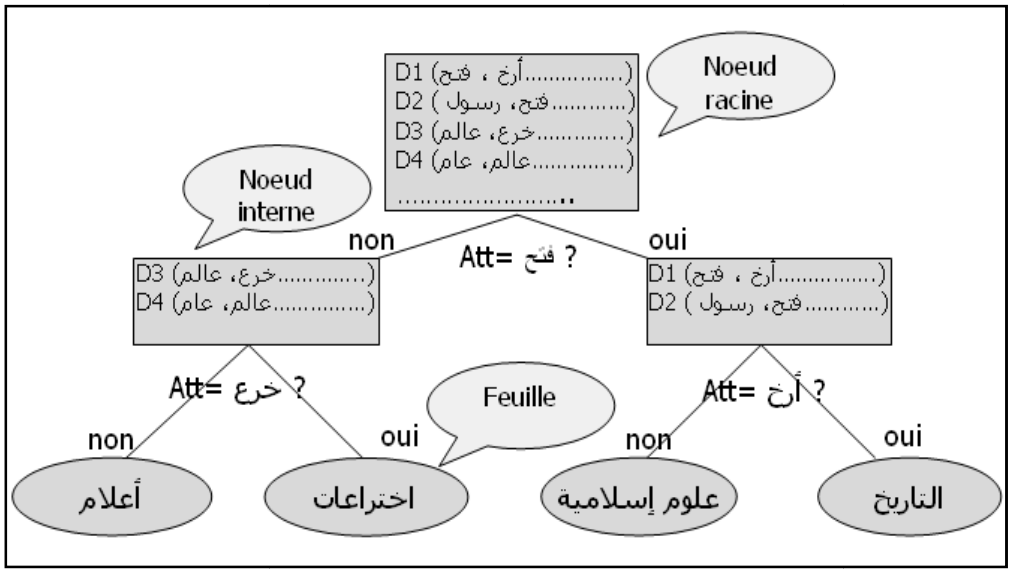

Fig. 5.1. Un exemple de l'arbre de décision résultant de l'opération de division des attributs

Les règles de classification sont directement déduites lors du parcours des différents chemins de l'arbre de décision, ainsi on a :

Si "*Fataha*, فتح " = "oui" *et* "*Arrakha*, أرّخ " = "oui" *alors la classe est* : Historique.

Si "*Fataha*, فتح " = "oui" *et* "*Arrakha*, أرّخ " = "non" *alors la classe est* : Sciences Islamiques.

Si "*Fataha*, فتح " = "non" *et* " K*harâa*, اخترع " = "oui" *alors la classe est* : Invention.

Si "*Fataha*, فتح " = "non" *et* " K*harâa*, اخترع " = "non" *alors la classe est* : Hommes célèbres.

Ces règles seront par la suite utilisées dans la phase de classification comme il est mentionné da la Section 2. Les documents de l'ensemble de test seront traités de la même manière que ceux de l'ensemble d'apprentissage, chaque document traité sera classé par l'ensemble de règles qui détermineront sa classe en fonction des attributs qu'il contient. Les résultats obtenus seront ensuite utilisés pour évaluer les performances du classificateur.

3. Préparation des corpus

Nous allons tout d'abord présenter les caractéristiques générales des corpus utilisés, pour mieux évaluer notre classificateur nous avons opté pour l'utilisation de deux corpus différents en langue arabe, le premier corpus qu'on a appelé "*Corpus hadith*" est un ensemble de traditions prophétiques (Hadiths du prophète '*qsassl*') qui se caractérise par l'éloquence de sa langue arabe et la spécialisation de son domaine, le deuxième corpus qu'on a appelé "*Corpus Scientifique*" est un ensemble de textes arabes générales de différents domaines.

3.1. Corpus Hadith

L'ensemble des documents de ce corpus se constitue d'une série de traditions prophétiques extraite à partir du CDROM de l'encyclopédie (*Alkotob Altissâa* الكتب التسعة) [Harf, 1997]. Les documents de cet ensemble sont distribués selon 14 catégories, par la moyenne de 25 textes par catégorie, données dans le Tableau 5.1.

La Catégorie	Nombre de
La Foi (الإيمان)	23
Le Coran (القرآن)	24
Le Savoir (العلم)	22
Les Crimes (الجنايات)	22
Al Djihad (الجهاد)	24
La Morale (الأخلاق والآداب)	31
Les Générations antérieures (الأمم السابقة)	12
La Bibliographie (السيرة)	11
Les Jugements (الأقضية والأحكام)	24
Les Adorations (العبادات)	23
Les Comportements (المعاملات)	25
L'Alimentation (الأشربة و الأطعمة)	31
Les Vêtements (اللباس والزينة)	34
Les Etats personnelles (الأحوال الشخصية)	24

Tableau 5.1. Descriptif des catégories du corpus Hadith

L'encyclopédie (*Alkotob Altissâa* الكتب التسعة) a été réalisée par la compagnie *SAKHR*, qui travaille déjà sur la création d'un classificateur thématique des traditions prophétique et du coran, nous avons retenue la même liste des catégories thématiques qui a été utilisée dans ce logiciel citée précédemment dans le Tableau 1. Le CD contient environ 62279 traditions prophétiques parmi les quelles nous avons choisi un échantillon de 340 traditions prophétiques comme documents de la phase d'apprentissage et 113 traditions prophétiques comme documents de la phase de test ce qui représente un total de 453 documents pour ce premier corpus.

3.2. Corpus scientifique

La collection des documents de ce corpus est extraite du CD-ROM de l'encyclopédie scientifique arabe "*Hal Tâalam* هل تعلم؟" [Arriss, 2001] qui contient environ 1092 documents parmi lesquelles nous avons choisi 280 documents pour l'ensemble d'apprentissage et 93 documents pour l'ensemble de test distribués sur 8 catégories, par moyenne de 35 documents pour chaque catégorie, données dans le Tableau 5.2.

La Catégorie	Nombre de Documents
Invention (اختراعات)	35
Géographie (جغرافيا)	35
Sport (رياضة)	35
Les hommes célèbres (أعلام)	35
Science Islamique (علوم إسلامية)	35
Histoire (التاريخ)	35
Science du Corps Humain (جسم	35
Cosmologie (علم الفضاء)	35

Tableau 5.2. Descriptif des catégories du corpus scientifique

4. Méthodologie de recherche

L'objectif de cette étude est l'évaluation du classificateur Arbre de Décision pour deux corpus en langue arabe. Les textes utilisés nécessitent le passage par les différentes étapes citées dans le Chapitre 3, Section 3. Dans la première étape, ces textes seront prétraités (nettoyage, élimination des mots vides et lemmatisation). Dans la deuxième étape le processus d'indexation sera appliqué sur l'ensemble des textes pour générer la matrice initiale qui sera divisée en deux ensembles: apprentissage et test. Dans la troisième étape, l'ensemble d'apprentissage sera utilisé pour la construction de l'arbre de décision selon l'algorithme discuté dans la section 2. Dans la quatrième étape on procède à l'extraction des règles de classification à partir de l'arbre construit. Dans la dernière étape, les textes de l'ensemble de test seront classés par ses règles pour mesurer et évaluer les performances de notre classificateur.

4.1. Expériences et résultats

La première étape de ces expériences est l'extraction du lexique de l'ensemble des textes de deux corpus. Nous avons d'abord supprimé les mots vides du lexique initial. Pour ce faire, nous avons employé une liste de mots vides de la langue arabe. Nous avons aussi appliqué un processus de lemmatisation[1] sur ce lexique. Notre processus de lemmatisation ne s'appuie pas sur des règles de dépendances syntaxiques mais sur des règles morphologiques. Nous disposons d'un dictionnaire de lemmes en langue arabe. Soit le mot appartient au dictionnaire soit le système nous retourne le mot lui-même comme lemme.

4.1.1. Expérience (1) :"Sélection d'attributs"

Cette expérience consiste à évaluer l'influence du choix de critère de sélection d'attributs sur les performances du système de classification. Les lexiques extraits à partir des deux corpus seront soumis à plusieurs opérations de filtrage et de sélection visant à diminuer substantiellement la taille, ainsi plusieurs critères peuvent être utilisés dans ces opérations. Nous avons choisi de tester les deux critères les plus utilisés dans le domaine de la réduction du vocabulaire à savoir:

- *La fréquence du terme (**TF**)*: Représente le nombre d'occurrences de terme dans un document.
- *La fréquence du document (**DF**)* : Représente le nombre de documents qui contiennent ce terme.

Généralement, ces critères peuvent être utilisés seuls, ou bien combinés, nous avons défini alors :

- *La fréquence combinée (**TF/DF**)* : qui Représente le nombre d'occurrences de terme dans un document combiné avec le nombre de documents qui contiennent ce terme.

1 On a utilisé le programme Al-Stem de *K. Darwish*, http://www.glue.umd.edu/~kareem/research/.

a. Corpus Hadith

Le corpus Hadith est composé de 340 documents, la taille du lexique retenue après l'application du critère de sélection est de 1938 termes. Les résultats de la classification des textes de ce corpus sans l'utilisation d'une technique de sélection d'attributs sont illustrés dans le Tableau 5.3.

Catégorie	Précision	Rappel	Erreur	Exactitude	Mesure F1
La Foi	0.33	0.11	0.08	0.91	0.16
Le Coran	0.25	0.11	0.09	0.90	0.15
Le Savoir	0.10	0.11	0.14	0.85	0.10
Les Crimes	0.14	0.11	0.11	0.88	0.12
Al Djihad	0.37	0.33	0.09	0.90	0.35
La Morale	0.80	0.50	0.04	0.95	0.61
Les Générations antérieures	0.12	0.25	0.17	0.83	0.16
La Bibliographie	0.33	0.50	0.10	0.89	0.40
Les Jugements	0.14	0.37	0.20	0.79	0.20
Les Adorations	0.67	0.75	0.04	0.95	0.70
Les Comportements	0.25	0.12	0.08	0.92	0.16
L'Alimentation	0.11	0.12	0.12	0.87	0.11
Les Vêtements	1.0	0.17	0.04	0.96	0.28
Les Etats personnelles	0.33	0.17	0.06	0.94	0.22
La moyenne	**0.35**	**0.27**	**0.01**	**0.89**	**0.26**

Tableau 5.3. Résultats de classification du corpus Hadith sans utilisation de la sélection d'attributs (Taille= 340 documents)

Pour la classification des textes du corpus Hadith avec l'utilisation d'une technique de sélection d'attributs, plusieurs tests ont été fait pour le choix du meilleur seuil pour chacun de ces critères et les résultats sont illustrés dans la Figure 5.2 et le Tableau 5.4.

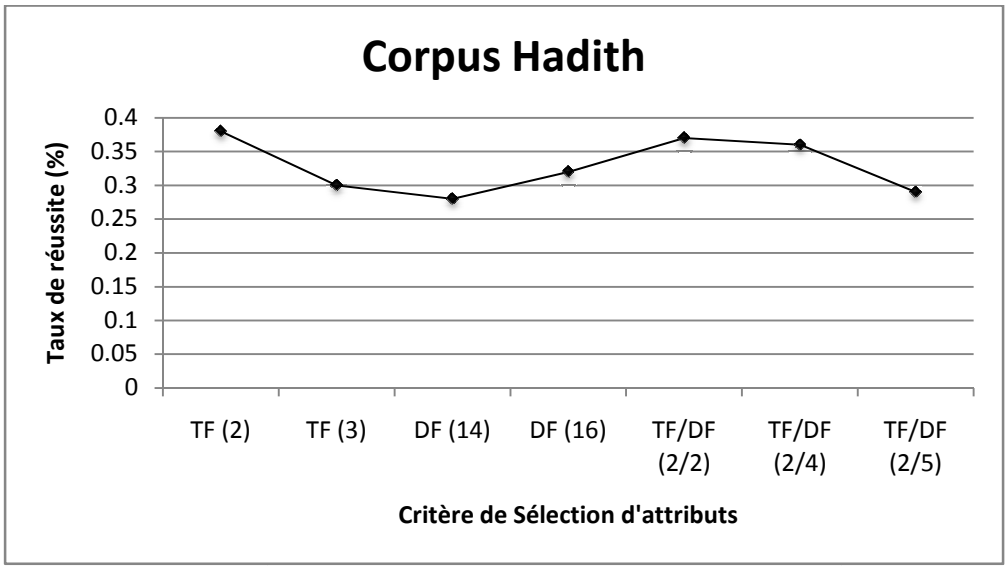

Fig. 5.2. Choix du meilleur seuil de sélection d'attributs pour le corpus Hadith.

Critère	N	P	E	Pourcentage (%)
TF (2)	113	43	70	**38%**
TF (3)	113	35	78	30%
DF (14)	113	32	81	28%
DF (16)	113	37	76	32%
TF/DF (2/2)	113	42	71	**37%**
TF/DF (2/4)	113	41	72	**36%**
TF/DF (2/5)	113	33	82	29%

Tableau 5.4. Choix du seuil de la fréquence pour le corpus Hadith

Où *N* représente le nombre total de document du corpus de test, *P* représente le nombre de documents bien classés et *E* représente le nombre de documents mal-classés. On constate que l'utilisation du critère de la fréquence des termes (*TF*) pour un seuil égal à deux (**2**) donne les meilleurs pourcentages. Pour le critère *TF* = 2 utilisé seul le taux de réussite du système est égal à 38 (%) et pour le critère *TF*=2 combiné avec le critère *DF*=2 le taux et égal à 37 (%).

b. Corpus Scientifique

Le corpus scientifique est composé de 280 documents, la taille du lexique retenue après l'application du critère de sélection est de 1107 termes. Les résultats de la classification des textes pour ce corpus sans l'utilisation de critère de sélection d'attributs sont illustrés dans le Tableau 5.5.

Catégorie	Précision	Rappel	Erreur	Exactitude	Mesure F1
Invention	0.50	0.17	0.12	0.88	0.25
Géographie	0.37	0.75	0.19	0.81	0.50
Sport	0.41	0.58	0.15	0.85	0.48
Les hommes célèbres	0.28	0.22	0.12	0.88	0.25
Sciences Islamiques	0.50	0.67	0.12	0.88	0.57
Histoire	0.28	0.17	0.14	0.85	0.21
Science du Corps	0.71	0.42	0.09	0.91	0.53
Cosmologie	0.64	0.58	0.09	0.91	0.61
La moyenne	**0.46**	**0.44**	**0.13**	**0.87**	**0.42**

Tableau 5.5. Résultats de classification du corpus Scientifique sans utilisation de la sélection d'attributs (Taille= 280 documents)

Les tests de choix du meilleur seuil pour les critères de sélection d'attributs sont illustrés dans le Tableau 5.6 et la Figure 5.3.

Critère	N	P	E	Pourcentage
TF (2)	9	5	4	53%
TF (3)	9	6	2	70%
TF (4)	9	5	4	53%
DF (8)	9	4	4	52%
DF (10)	9	4	4	50%
DF (11)	9	5	4	53%
TF/DF	9	4	4	47%
TF/DF	9	4	4	48%
TF/DF	9	4	5	43%
TF/DF	9	4	4	48%
TF/DF	9	2	6	27%

Tableau 5.6. Choix du seuil de la fréquence pour le corpus scientifique

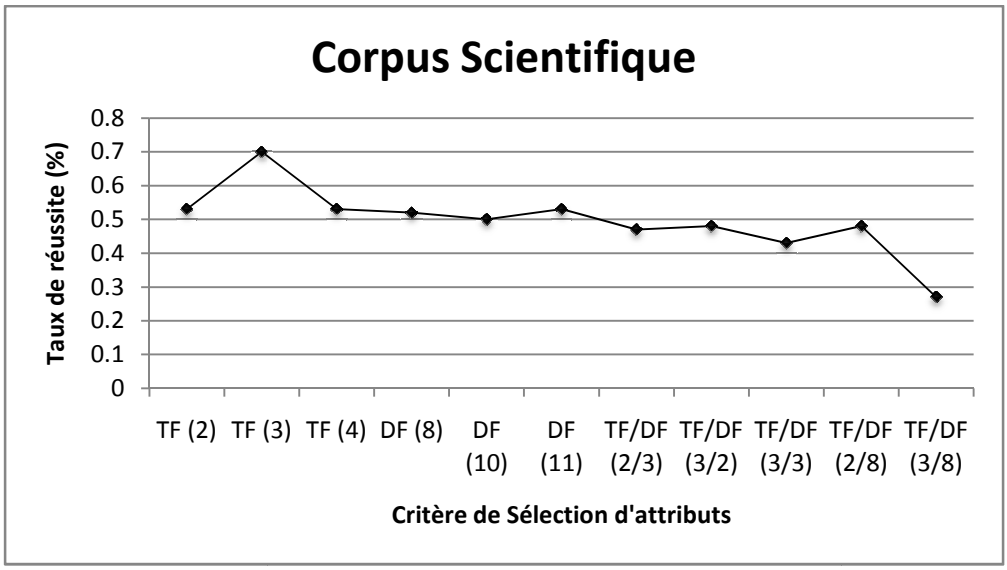

Fig. 5.3. Choix du meilleur seuil de sélection d'attributs pour le corpus Scientifique

D'après ces résultats on constate que l'utilisation du critère de la fréquence des termes (TF) pour un seuil égal à trois (3) donne le meilleur pourcentage 70(%).

c. Discussions des résultats de l'expérience (1)

La comparaison des résultats de classification avec et sans utilisation du critère de sélection d'attributs pour les deux corpus est illustrée dans les figures 5.4 et 5.5. Nous avons fixé le seuil de la fréquence des termes à 2 pour le corpus littéraire et à 3 pour le corpus scientifique.

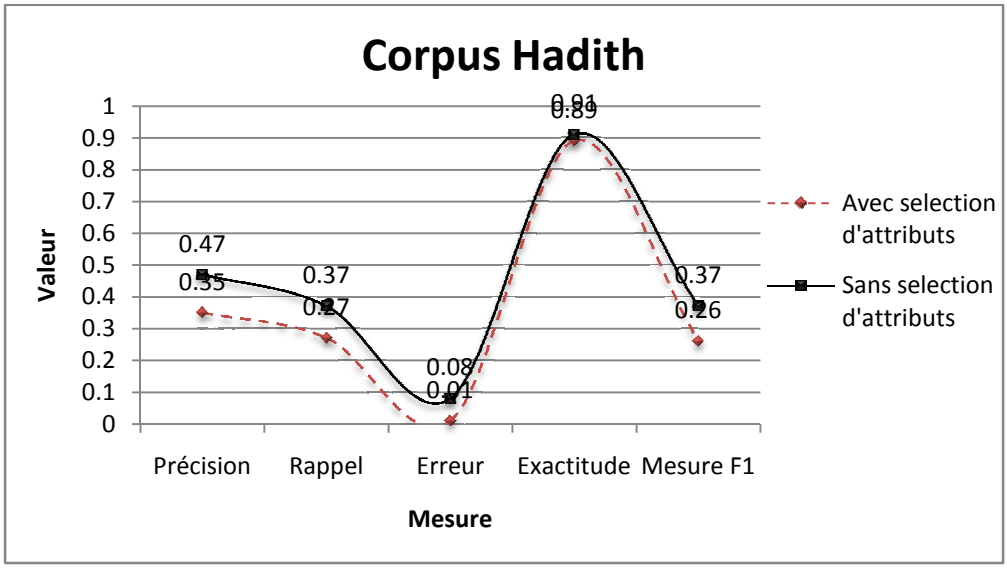

Fig. 5.4. Influence du critère « sélection d'attributs » sur les performances du classificateur pour le corpus Hadith (340 documents)

D'après la figure 4.4, on constate une amélioration de 12(%) pour la précision, de 10(%) pour le rappel, de 7(%) pou le taux d'erreur, de 2(%) pour le taux d'exactitude et de 11(%) pour la mesure F1. L'amélioration globale de performance pour le corpus Hadith est autours de 11(%).

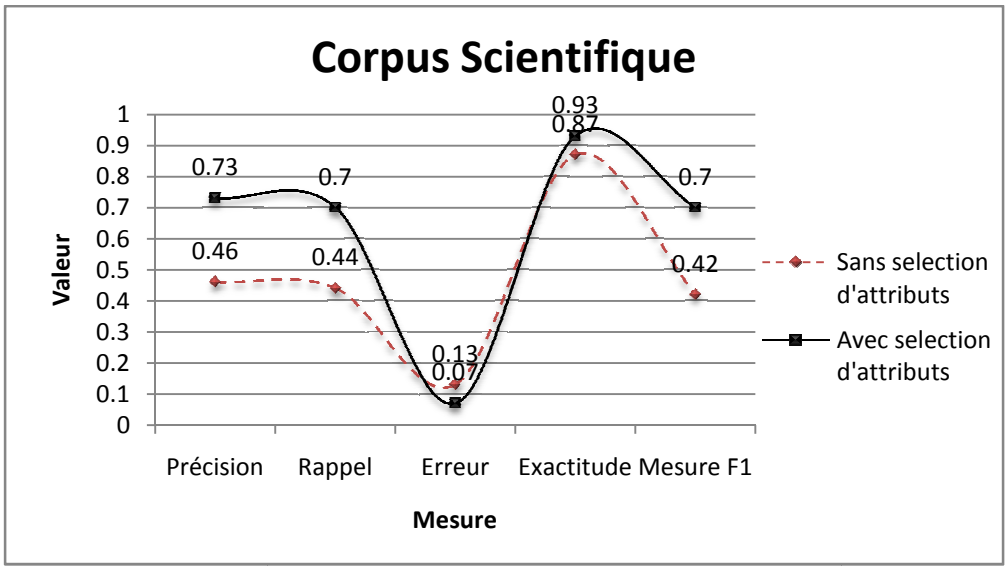

Fig. 5.5. Influence des critères de sélection d'attributs sur les performances du classificateur pour un corpus scientifique de 280 documents

D'après la figure 4.5, on constate une amélioration de 27(%) pour la précision, de 26(%) pour le rappel, de 6(%) pou le taux d'erreur, de 6(%) pour le taux d'exactitude et de 28(%) pour la mesure F1. L'amélioration globale de performance pour le corpus scientifique est autours de 26.5 (%).

Les résultats de cette expérience confirment que pour la phase de réduction de vocabulaire, la combinaison d'un critère de sélection d'attributs avec le critère du gain d'information, déjà existant dans l'algorithme des arbres de décision lui-même, permet d'améliorer considérablement les résultats de classification.

4.1.2. Expérience (2) :"Taille de l'ensemble d'apprentissage"

La deuxième expérience consiste à évaluer l'influence de la taille de l'ensemble d'apprentissage sur les performances du système de classification. Nous avons utilisé les résultats de la première expérience tout en fixant le seuil de la fréquence des termes à 2 pour le corpus Hadith et à 3 pour le corpus scientifique.

a. Corpus Hadith

Les Figures 5.6, 5.7, 5.8, 5.9 et 5.10 montrent les résultats des tests obtenus pour le corpus Hadith, les différents tailles de l'ensemble d'apprentissage sont respectivement : 150, 191, 239, 266, 340.

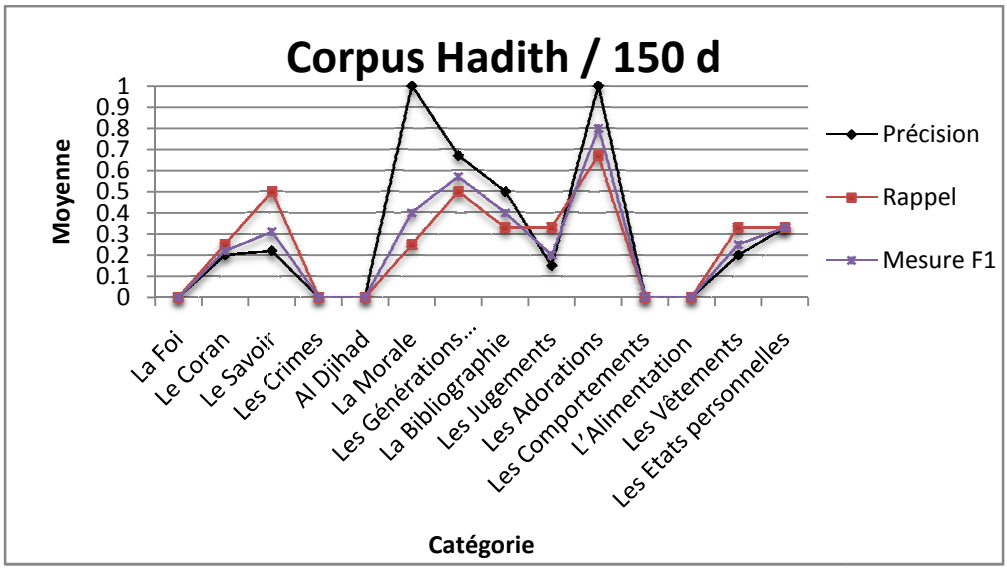

Fig. 5.6. Les résultats de classification du corpus Hadith pour un ensemble d'apprentissage de "150" documents.

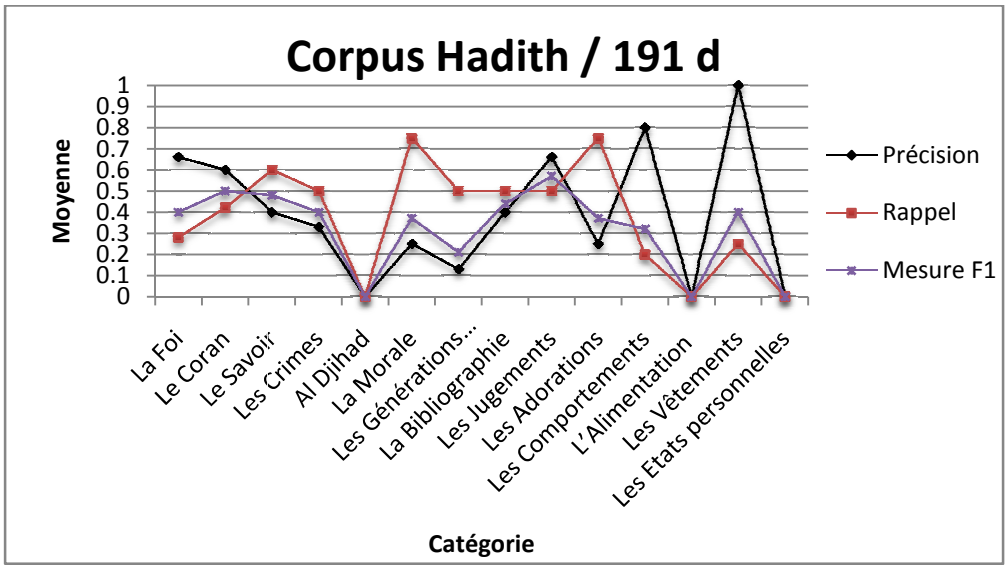

Fig. 5.7. Les résultats de classification du corpus Hadith pour un ensemble d'apprentissage de "191" documents.

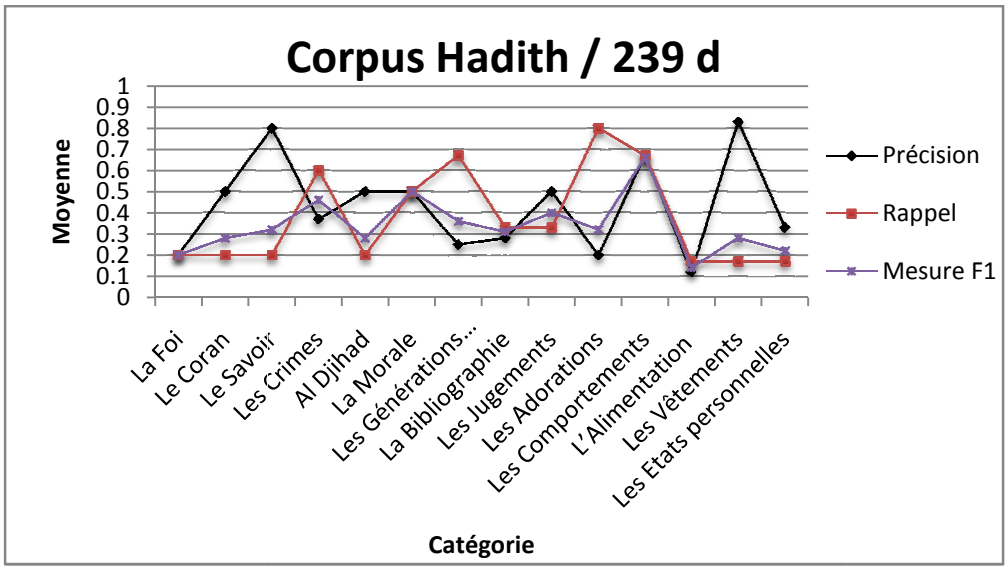

Fig. 5.8. Les résultats de classification du corpus Hadith pour un ensemble d'apprentissage de "239" documents.

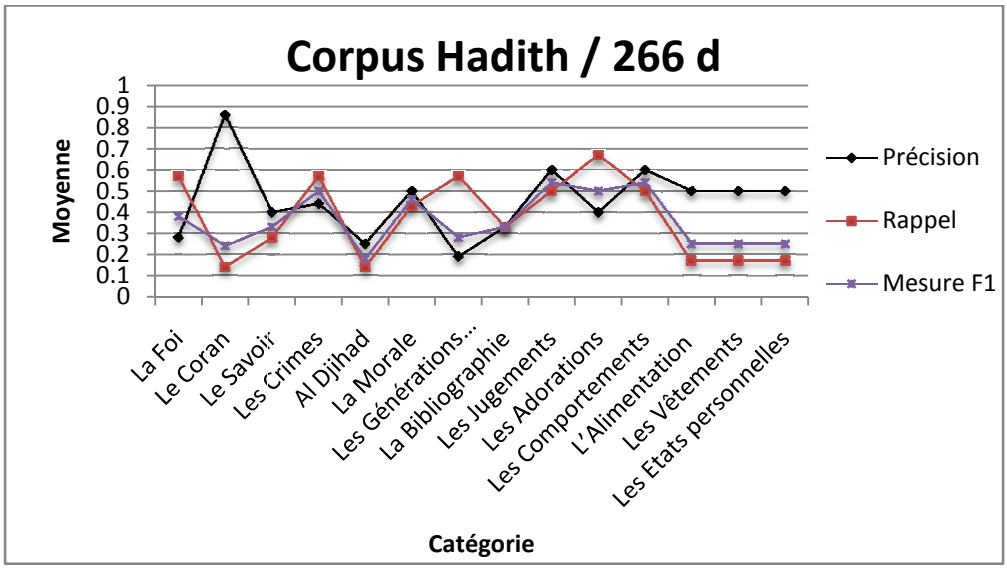

Fig. 5.9. Les résultats de classification du corpus Hadith pour un ensemble d'apprentissage de "266" documents.

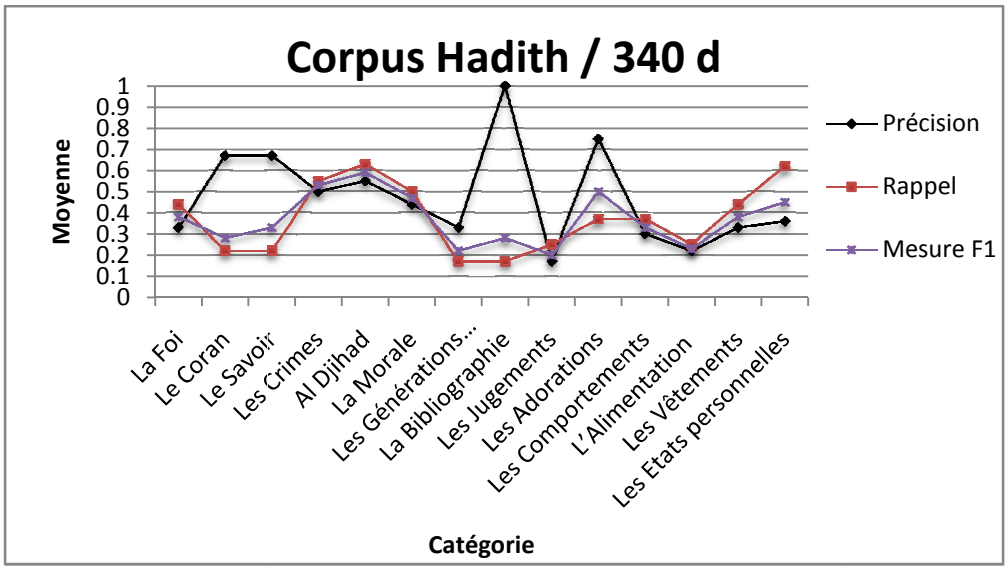

Fig. 5.10. Les résultats de classification du corpus Hadith pour un ensemble d'apprentissage de "340" documents.

b. Corpus scientifique

Les Figures 5.11, 5.12, 5.13, 5.14 et 5.15 montrent les résultats des tests obtenus pour le corpus scientifique, les différents tailles de l'ensemble d'apprentissage sont respectivement : 100, 140, 180, 200, 280.

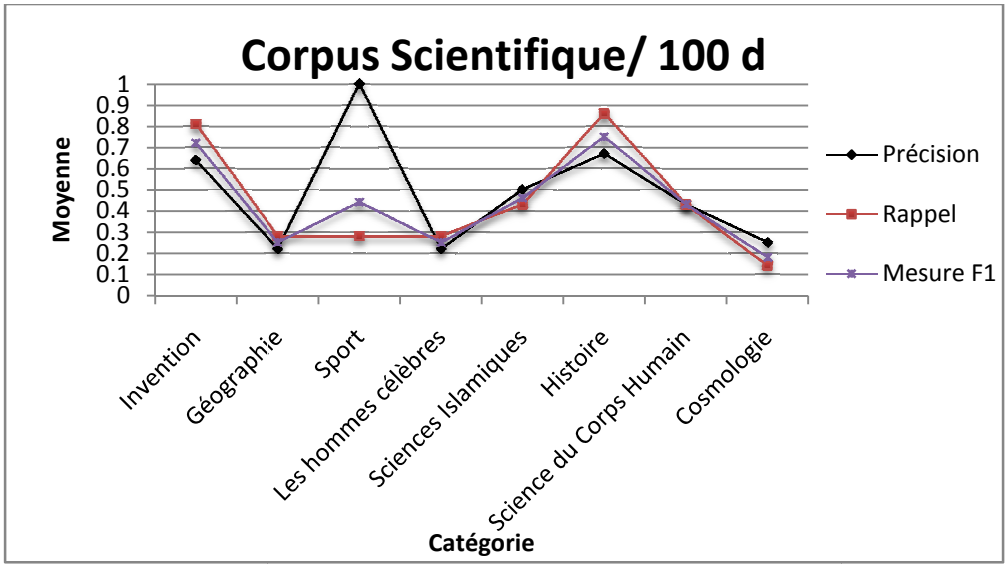

Fig. 5.11. Les résultats de classification du corpus Scientifique pour un ensemble d'apprentissage de "100" documents.

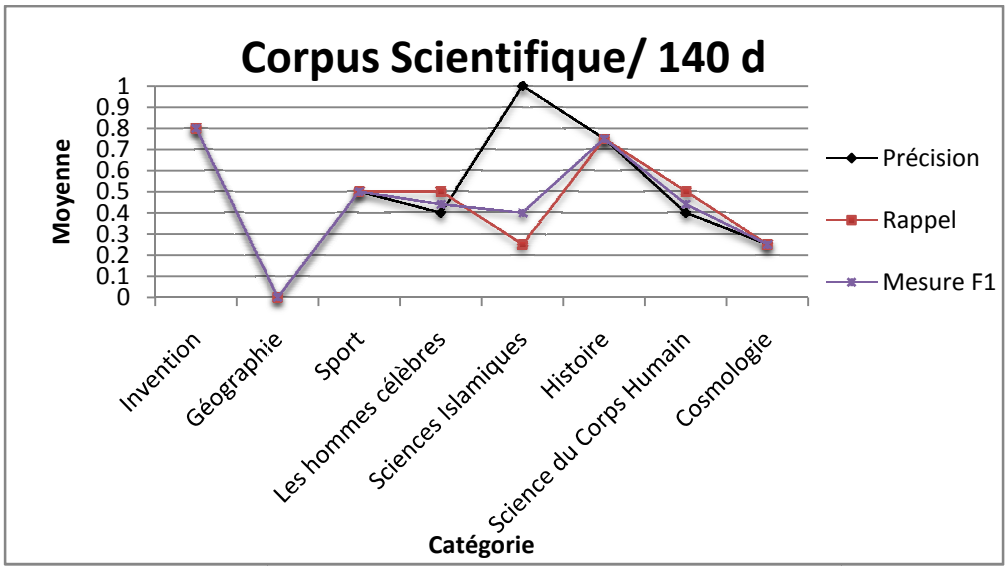

Fig. 5.12. Les résultats de classification du corpus Scientifique pour un ensemble d'apprentissage de "140" documents.

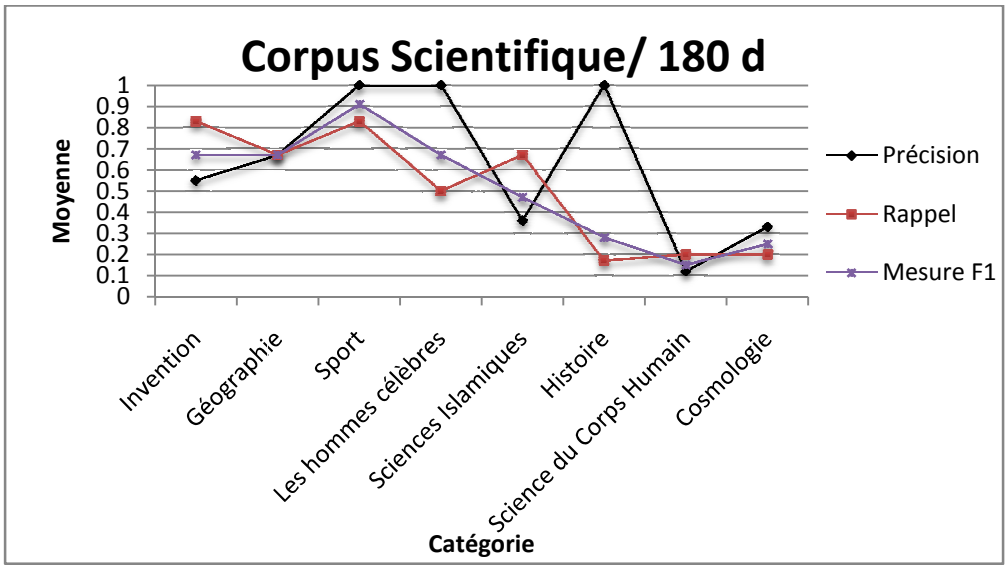

Fig. 5.13. Les résultats de classification du corpus Scientifique pour un ensemble d'apprentissage de "180" documents.

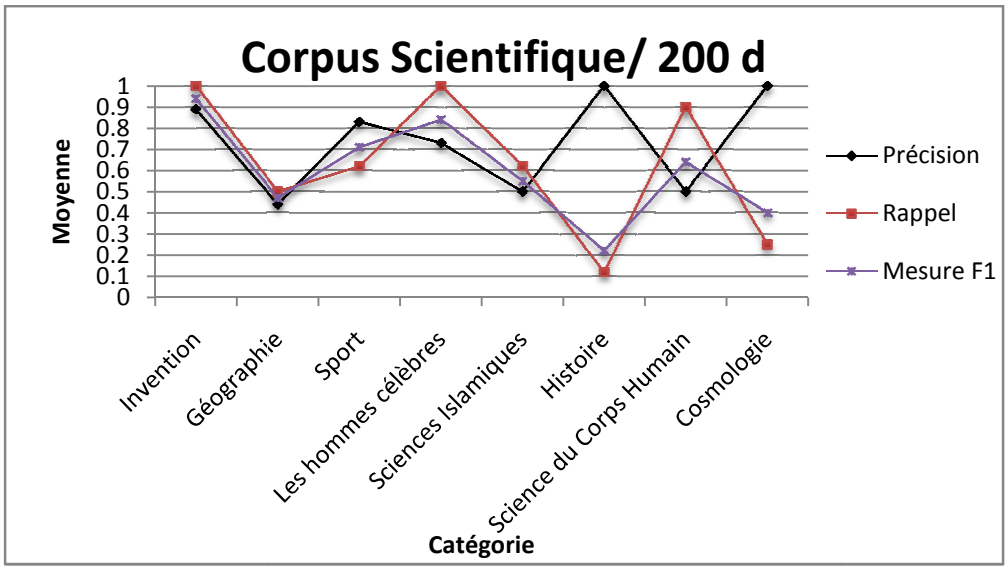

Fig. 5.14. Les résultats de classification du corpus Scientifique pour un ensemble d'apprentissage de "200" documents.

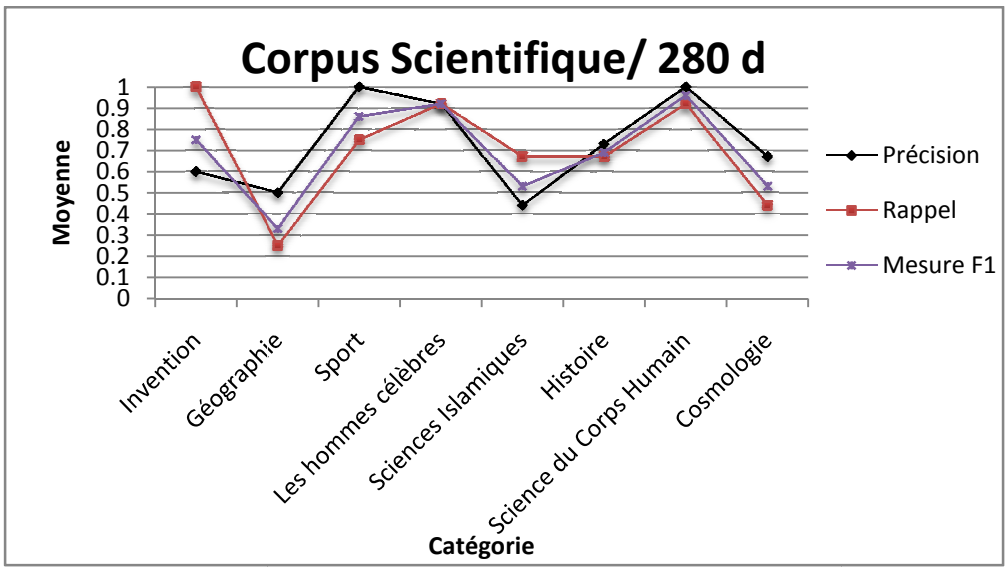

Fig. 5.15. Les résultats de classification du corpus Scientifique pour un ensemble d'apprentissage de "280" documents.

b. Discussions des résultats de l'expérience (2)

D'après la figure 5.16, on constate qu'entre la taille de 150 documents et celle de 340 documents il y a une amélioration de 17(%) pour la mesure de précision et de 13(%) pour la mesure F1. Il y a presque une stabilité pour les valeurs des trois autres mesures 0.37 pour le rappel, de 0.08 pour le taux d'erreur, de 0.91 pour le taux d'exactitude, ce qui donne une amélioration globale de performance autours de 15(%) pour le corpus Hadith.

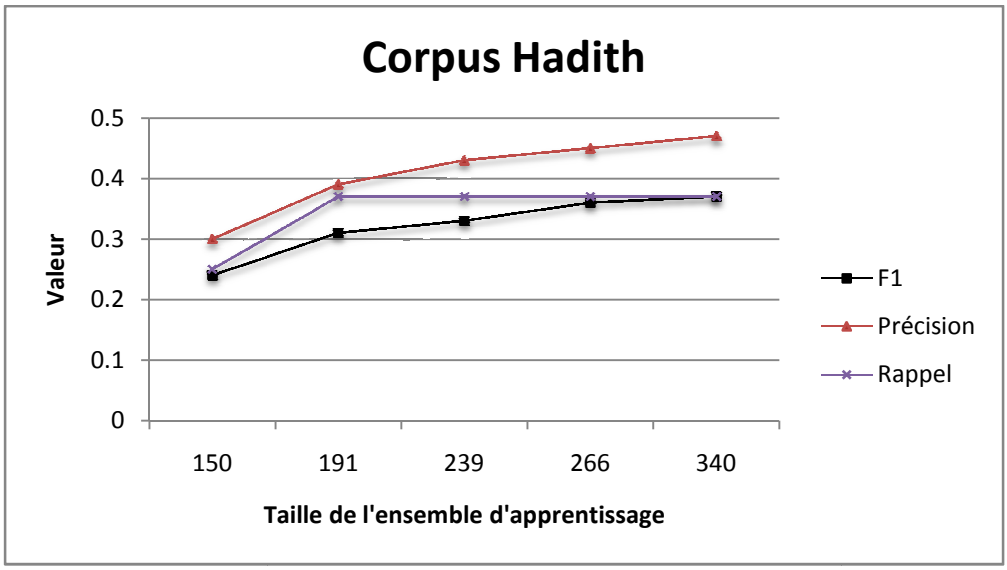

Fig. 5.16. Influence du critère de la taille d'ensemble d'apprentissage sur les performances du classificateur pour le corpus Hadith.

D'après la figure 5.17, on constate qu'entre la taille de 100 documents et celle de 280 documents il y a une amélioration de 24(%) pour la mesure de précision, de 26(%) pour la mesure de rappel et de 27(%) pour la mesure F1, et il y a presque une stabilité pour les valeurs des deux autres mesures 0.09 pour le taux d'erreur et de 0.90 pour le taux d'exactitude. Donc pour le corpus scientifique il y a une amélioration globale de performance autours de 26.5 (%).

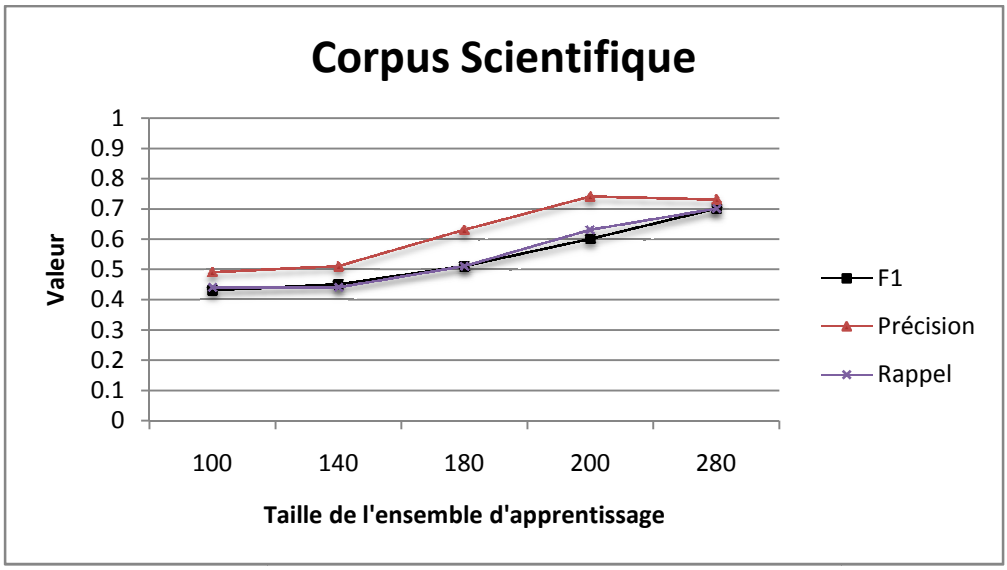

Fig. 5.17. Influence du critère de la taille d'ensemble d'apprentissage sur les performances du classificateur pour le corpus Scientifique.

Les résultats de cette expérience confirment que l'augmentation de la taille de l'ensemble d'apprentissage à une grande Influence sur les performances du classificateur.

4.1.3. Discussion des résultats de classification

La Figure 5.18 montre les résultats moyens des mesures d'évaluation : Précision, Rappel et Exactitude pour chaque catégorie (classe) du corpus Hadith. Nous pouvons dire que la classe "*Les vêtements*" a eu la meilleure valeur pour la mesure de précision (0.57) suivie de la classe "*Le Coran*" pour une valeur de (0.53), la classe " *Les Adorations* " a eu la meilleure valeur pour la mesure de rappel (0.65) suivie par les deux classes "*La Morale*" et " *Les Générations antérieures*" pour une valeur de (0.48). Tandis que la meilleure valeur pour la mesure d'exactitude (0.94) a été attribuée à la classe "*Les Adorations*" suivie de la classe "*La Morale* " pour une valeur de (0.93).

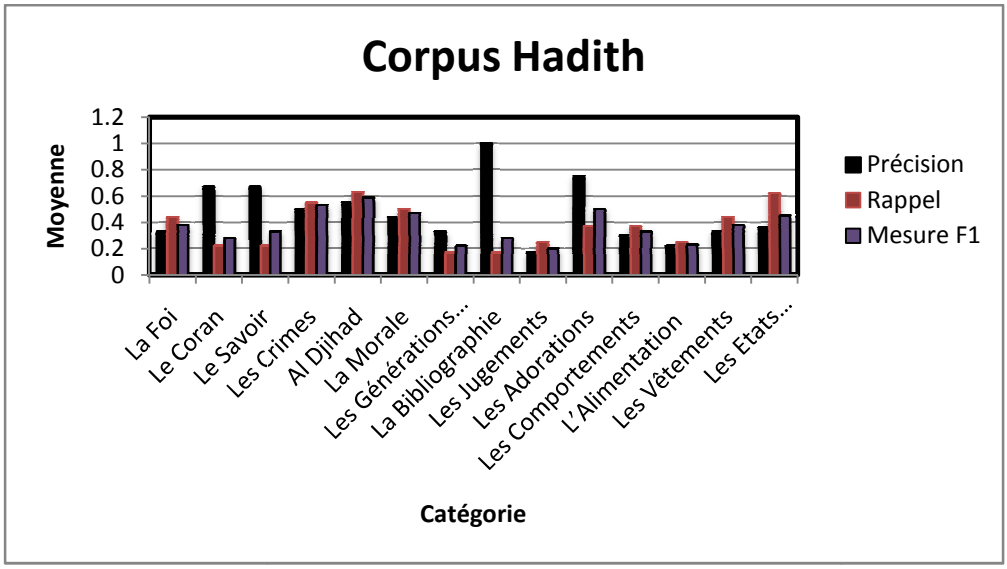

Fig. 5.18. Les résultats moyens des mesures d'évaluation pour les classes du corpus Hadith.

La Figure 5.19 montre les résultats moyens des mesures d'évaluation pour les catégories du corpus scientifique. D'après cette figure on constate que la meilleure valeur pour la mesure de précision (0.86) a été eut par la classe "*Sport*" suivi de la classe "*Histoire*" avec une valeur de (0,83). Pour la mesure de rappel la meilleure valeur (0.88) a été eut par la classe "*Invention*" suivi de la classe " *Les hommes célèbres* " avec une valeur de (0.64). Tandis que la meilleure valeur pour la mesure d'exactitude (0.93) a été eut par la classe " *Sport* " suivi par les deux classes " *Invention* " et "*Histoire*" pour une valeur de (0.92).

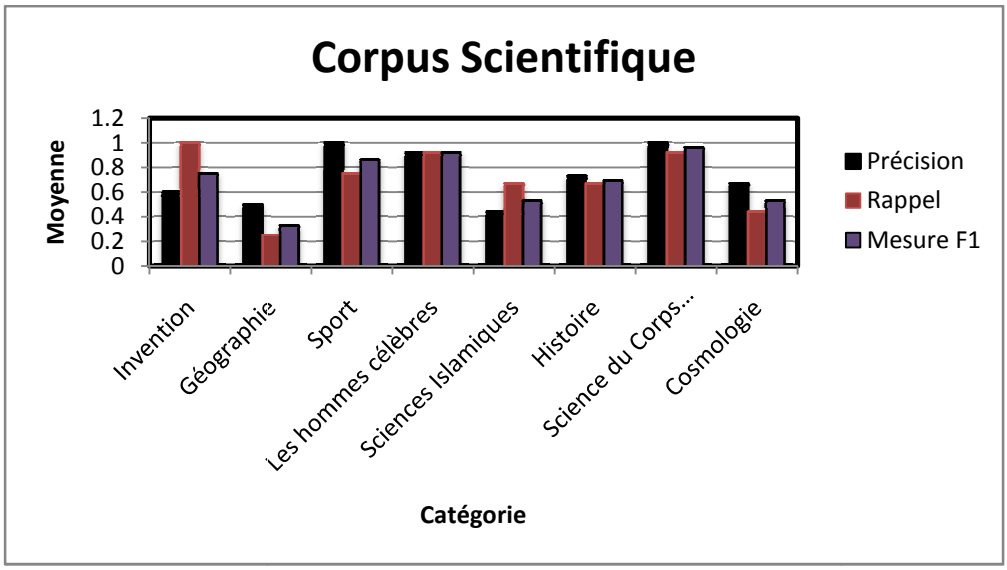

Fig. 5.19. Les résultats moyens des mesures d'évaluation pour les classes du corpus scientifique.

4.2. Comparaison des résultats d'évaluation des deux Corpus

En comparant les résultats d'évaluation obtenus à partir des différentes expériences effectués sur les deux corpus, les meilleurs résultats sont ceux des documents du corpus scientifique comme donnés par la Figure 5.20.

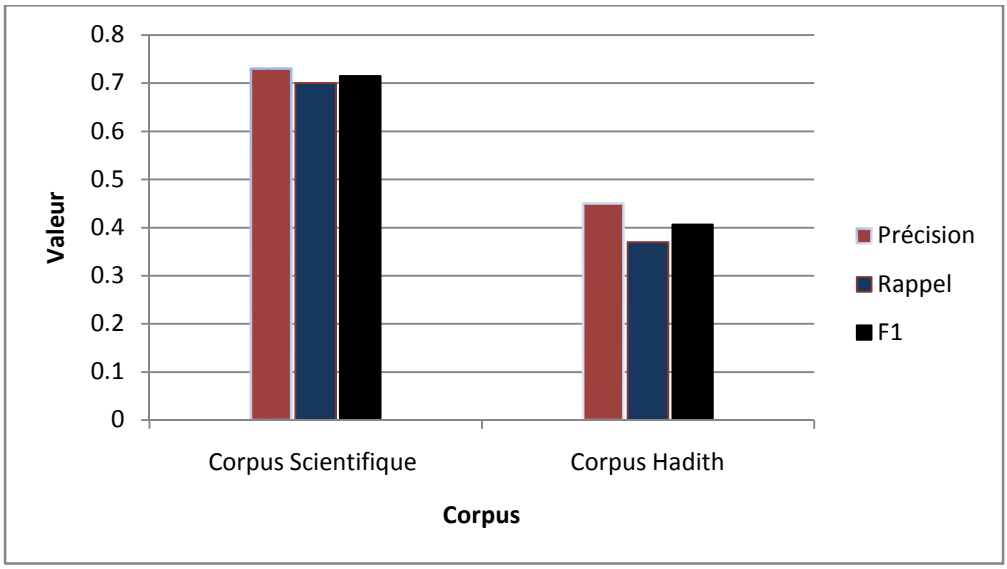

Fig. 5.20. Comparaison des résultats d'évaluation pour les deux corpus.

On voit bien que la valeur de la mesure F1 pour le corpus scientifique est supérieure à celle du corpus Hadith. Cette mesure nous donne une idée globale sur l'efficacité et la performance du classificateur pour chaque corpus, plus la valeur de cette mesure est proche de 100% plus les performances du classificateur en terme de précision et de rappel sont meilleurs.

La faiblesse des résultats de l'évaluation du corpus Hadith, nous a conduits à une révision des règles de classification extraites précédemment à partir de l'arbre de décision. Cette révision a mis en évidence que la plupart des erreurs de classification sont en vérité dû à la nature et aux caractéristiques des documents du corpus Hadith. La plupart de ces documents contiennent un

grand nombre de termes qui peuvent apparaîtraient dans plus d'une catégorie ou d'une classe, en d'autres termes, nous constatons qu'un bon nombre de documents qui sont susceptibles d'être classés dans une catégorie "*X*" contiennent beaucoup de termes qui ont une fréquence plus élevée dans d'autres catégories ce qui a une influence directe sur les résultats de la classification. Par exemple, nous constatons qu'il y a beaucoup d'erreurs dans la classification de documents de la catégorie "*La Foi*" dans de la catégorie "*Les Adorations*", et *vice-versa,* et ceci est dû principalement au fait que les documents des deux catégories contiennent un grand nombre de termes communs comme : "*Intention* النيّة" pour la catégorie " *Adorations*" et pour la catégorie "*La Foi*", "*pureté* الطهارة" pour "*la pureté du cœur*" et pour "*la pureté du corps*". La même remarque peut être faite pour les deux catégories "*Al djihad*" et "*Les Crimes*" car il y a un ensemble de mots communs tels que: "*Meurtre* القتل", "*Ame* النفس" et "*Déliement* العتق".... Etc. En plus des facteurs de sélection d'attributs et de la taille de l'ensemble d'apprentissage qui ont été abordés au cours des expériences précédentes, on peut aussi indiquer qu'il y a d'autres facteurs qui peuvent influer sur les performances du système de classification. Le premier facteur est celui de la divergence thématique entre les différentes catégories. Ce facteur peut nécessairement forcer l'existence d'une série de termes discriminatoires pour chaque catégorie, ce qui permettra d'améliorer les résultats de classification d'une similaire au corpus scientifique.

Le deuxième est celui de la taille du document lui-même (non-expérimenté) néanmoins nous avons constaté, lors de l'examen du contenu de certains documents du corpus littéraire (*Traditions Prophétiques*), qu'ils ont une petite taille comparée à celle des documents du corpus scientifique. Cette taille ne dépasse pas la moitié d'une page pour la plupart des cas, et dont il ne reste qu'un nombre restreint de termes pour représenter chaque document après l'application du processus de prétraitement, ce qui ne permet pas de donner une image claire sur sa classe.

4.3. Comparaison avec d'autres systèmes de classification

Dans un objectif d'évaluation de notre système de classification basé sur les arbres de décision, nous avons réalisé une comparaison avec d'autres systèmes de catégorisation des textes en langue arabe. Les résultats de cette comparaison sont exposés dans le Tableau 5.7. La comparaison a été effectuée avec des systèmes de classification probabilistiques basés sur l'algorithme bayésien naïf comme dans [El-Kourdi et al., 2004], des systèmes de classification statistiques basés sur l'algorithme du maximum entropie comme dans [Sawaf et al., 2001] et des systèmes de classification linéaires basés sur le modèle de l'espace vectoriel et utilisant les mesures de di-similarité de *Dice*, de *Jaccard* et la mesure *Euclidienne* comme dans [Al-Kabi & Al-Sinjilawi, 2007].

Cette comparaison montre que le système de classification basé sur l'algorithme des arbres de décision est l'un des meilleurs systèmes en termes de performance globale. On remarque que les systèmes de classification de bayésien naïf et de maximum entropie ont des valeurs supérieures aux notre pour la mesure de rappel néanmoins notre système présente de meilleures valeurs pour les mesures de F1 et de précision ce qui prouve que notre système et plus précis.

En termes de temps de classification, Les systèmes de classification basés sur les arbres de décision ont une bonne côte parmi les meilleurs systèmes. Nous avons procédé à la mesure du temps de classification des documents de l'ensemble de test pour les deux corpus (93 documents pour le corpus scientifiques) et (113 documents pour corpus littéraire). Ce temps ne dépasse pas les 120 secondes, très court par rapport au temps du système de classification basé sur l'algorithme K-PPV mentionné dans [Syiam et al. 2006], car ce système a atteint un temps de 3004 seconds soit environ 5,04 minutes. Le rapport 1/25 entre ces deux durées est dû principalement à la complexité algorithmique qui est considérée très faible dans le cas de l'algorithme des arbres de décision, elle est égale à *O(LogN)*, où *N* est le nombre de nœuds dans l'arbre.

Système	Précision	Rappel	Mesure F1
Arbres de decision	73.00	70.00	70.00
Bayésien naïf	67.88	71.96	67.83
Maximum entropie	50.00	84.20	62.70
Modèle de l'éspace vectoriel (Dice)	41.00	44.00	42.00
Modèle de l'éspace vectoriel (Jaccard)	54.00	61.00	57.00
Modèle de l'éspace vectoriel (Euclidiènne)	54.00	57.00	55.00

Tableau. 5.7. Comparaison des résultats d'évaluation du classificateur des arbres de décision avec d'autres systèmes de classification.

5. Conclusion

Dans ce chapitre, nous nous sommes intéressés à l'évaluation d'un système de classification basé sur l'algorithme des arbres de décision. Nous avons présenté et étudié plusieurs mesures et critères utilisés dans l'évaluation des performances des systèmes de classification. Notre recherche a été enrichie par plusieurs expériences qui ont permis d'étudier l'impact des critères de sélection d'attributs et de la taille de l'ensemble d'apprentissage sur les performances et l'efficacité du système de classification présenté. Pour une meilleure évaluation de notre système, nous avons opté pour l'utilisation de deux corpus différents en langue arabe (Hadith et corpus scientifique). Nous avons également analysé les différents résultats d'évaluation obtenus pour chaque corpus ainsi que les résultats de leur comparaison ce qui nous a permis de conclure qu'une série facteurs peuvent influencer le fonctionnement du système de classification en particulier la nature et la spécificité des documents de chaque corpus. Cette étude a démontré le caractère inévitable de l'utilisation des corpus constitués de classes les plus thématiquement divergentes et a aussi confirmé l'indépendance des méthodes statistiques (indexation, pondération, gain de l'information) et les méthodes de classification (algorithme des arbres de décision) *vis-à-vis* de la langue utilisée, ces méthodes restent valides pour la langue arabe comme pour d'autres langues Indo-Européennes.

Pour nos futurs travaux, nous envisageons d'effectuer plus d'expériences dans le but de trouver les meilleurs facteurs qui peuvent améliorer les performances de notre système de classification. Notre but est de passer de la classification mono-classe (*mono-thématique*) où le document peut être classé dans une seule catégorie à la classification multi-classe (*multi-thématique*) où le document peut être classé dans plus d'une catégorie avec l'introduction de techniques de la logique floue à cet algorithme. Cette technique est déjà connue sous le nom des arbres de décision floues, ou en utilisant les techniques de segmentation thématique des textes. L'utilisation de ces deux techniques va permettre de réduire le taux des erreurs de classification. La dernière perspective est la possibilité d'utiliser des moyens linguistiques tel que les extracteurs de racines, les analyseurs morphosyntaxiques et l'utilisation des ressource externes telles que les dictionnaires terminologies de la langue arabe ou les thésaurus et de mener d'autres expériences pour étudier l'impact de ces ressources sur les performances des systèmes de classification.