Chapitre 4

Evaluation des classifieurs

Table des matières

4.1- Introduction

Différentes approches décrites dans le chapitre 3, ont été utilisées pour la catégorisation de textes offrant ainsi, aux développeurs dans le domaine plusieurs issues, qui amène à poser une question très récurrente sur le choix du meilleur algorithme pour la classification automatique de textes.
Pour pouvoir répondre à cette question, il faut bien disposer de critères et tests utiles pour mesurer et évaluer les performances d'un classifieur pour pouvoir les comparer par la suite, afin, éventuellement, opter pour un classifieur ou un autre.

Mais qu'est ce que l'***évaluation*** ? L'évaluation consiste à mesurer la différence entre un résultat attendu et un résultat obtenu. Aucune métrique n'est associée, mais en général, on utilise des indicateurs compris entre 0 et 1 pour en faciliter l'interprétation. (Nakache & Metais, 2005).

Néanmoins, l'évaluation des performances d'un système de classification n'est pas toujours triviale et elle dépend de l'utilisation finale de ce système, certains éléments sont très subjectifs et difficilement automatisables.
Notons qu'il n'est pas facile de juger un système de catégorisation de textes s'il est performant ou moins performant qu'un autre. Plusieurs facteurs entrent en jeu qui rend cette évaluation relative, de la base textuelle à classer, de l'approche adoptée pour représenter les textes, de l'algorithme d'apprentissage opérée, et enfin du juge humain puisque l'attribution finale d'un document à une classe dépend du centre d'intérêt des utilisateurs de ses systèmes.

Ainsi, des mesures différentes, pour des systèmes dédiés à la classification déterministe ou « dure » et pour des systèmes dédiés à la classification floue ou ranking, sont proposées, qui s'intéressent chacune d'elles à un aspect de classification.
Dans ce chapitre nous allons présenter les méthodologies existantes pour aborder une vraie comparaison entre les classifieurs, pour ensuite dévoiler les mesures de performance souvent utilisées dans la littérature, et enfin achever par un description brève d'autres critères de performances non mesurables.

4.2- Méthodologies de comparaison de classifieurs

Il existe, en pratique, plusieurs méthodologies pour tenter de répondre à la question : quel est la meilleure méthode pour la catégorisation de textes ?

4.2.1- Différentes approches sur le même corpus

La première solution consiste à comparer différentes méthodes mises en œuvre par différents auteurs sur le même corpus, néanmoins, du point de vue pratique, comme le confirme Radwan JALAM dans (Jalam, 2003), on est confronté à pas mal de problèmes, parmi lesquels :

4.2.1.1- Même corpus avec des découpages différents

Les différents auteurs n'utilisent pas exactement le même découpage du corpus, par exemple pour Reuters seulement, il y a plus de six versions différentes, qui se distinguent par le nombre de leurs classes et la répartition des documents sur le corpus d'apprentissage et le corpus de test. Pour Reuters-21578 qui est souvent utilisé, (Joachims, 1998), (Schapire & all, 1998), (Yang & Liu, 1999) considèrent 90 catégories, (Dumais & all, 1998) en considèrent 118, d'autres travaillent carrément sur Reuters-top10 comme dans (Turenne, 2000) ou (Denoyer, 2004) ou (Yvon, 2006), qui trient les dix meilleurs catégories (Mini corpus utilisé dans nos expérimentations). De plus, la plupart des auteurs considèrent 3299 documents sur la

base de test, mais (Yang & Liu, 1999) en considèrent uniquement 3019 en supprimant tous les documents de la base de test qui n'appartiennent à aucune catégorie. Finalement, ces légères différences de découpage rendent difficiles les comparaisons à travers ces publications.

4.2.1.2- Les différentes techniques de représentation de textes

Les différentes alternatives offertes, pour le choix de descripteurs, afin de coder un texte, ainsi que les diverses méthodes de réduction de dimensionnalité utilisées par les différents auteurs peuvent embrouiller la comparaison de deux classifieurs s'exerçant sur le même corpus.

4.2.1.3- Les différentes mesures utilisées pour l'évaluation

Les mesures de performance utilisées dans les différentes expérimentations ne sont pas les mêmes (une description de quelques mesures est présentée dans la section suivante), ainsi les différents critères de performance peuvent être estimés de différentes façons empêchant une comparaison efficace entre les classifieurs.

4.2.2- Différentes approches par le même auteur

Une autre approche, plus crédible de point de vue scientifique, souvent proposée est l'utilisation de plusieurs méthodes par le même auteur, et automatiquement le corpus, le découpage de ce dernier, les techniques de codage, et les mesures de performance sont semblables pour toutes les méthodes. (Yang & Liu, 1999) comparent ainsi les kPPv, les SVM, les réseaux de neurones, et d'autres approches.
(Dumais & all, 1998) proposent également plusieurs comparaisons en mettant en opposition Les SVM, l'algorithme de Rocchio, les arbres de décision, et les réseaux bayesiens.

4.2.3- Difficultés approuvées pour juger les capacités d'une méthode

Les comparaisons présentées évaluent plus les compétences des auteurs dans l'exploitation des différentes approches de l'état de l'art les méthodes, plus que les capacités des méthodes elles-mêmes.
Le problème vient du fait que toutes ces méthodes sont délicates à mettre en œuvre et leurs performances dépendent fortement de leurs différentes utilisations.
Par exemple, l'implémentation des machines à vecteurs supports proposées par (Dumais & all, 1998) obtient nettement de meilleurs résultats que celle proposée par (Joachims, 1998).
Les réseaux de neurones testés par (Yang & Liu, 1999) sont des perceptrons multi-couches avec une couche cachée comportant 64 neurones, 1000 descripteurs en entrées et 90 neurones de sorties correspondant aux 90 catégories ; ils considèrent un seul réseau pour l'ensemble des catégories comportant plus de 64000 poids. Il n'est pas surprenant, dans ces conditions, que les performances obtenues ne soient pas très bonnes.
Il reste aussi difficile d'extrapoler les performances sur d'autres corpus et applications. Les résultats sont extrêmement dépendants du type des textes et des classes (en particulier de leur nombre). Il n'existe pas, à l'heure actuelle d'analyse de la performance des algorithmes en fonction des spécificités des corpus.
Ces différentes remarques prouvent que le succès d'une méthode dépend d'un ensemble de paramètres et certaines conditions non liées seulement, aux algorithmes d'apprentissage eux mêmes, mais aussi aux différents choix opérés pendant tout le processus, et qui peuvent intervenir et influencer les résultats obtenus. Par conséquent, il est extrêmement difficile de tirer des conclusions définitives sur une approche.

4.2.4- TREC

Il nous semble que la conférence TREC (Décrite en annexe) est une bonne solution pour comparer différentes méthodes, car chaque participant propose des solutions qu'il connaît bien avec des algorithmes dont il a pu tester l'efficacité. Le corpus est évidemment identique pour tout le monde, ainsi que les méthodes d'évaluation et la répétition annuelle de cette conférence permet de juger les approches sur le long terme.
De plus la conférence TREC a l'avantage de proposer un état de l'art à un instant donné contrairement aux comparaisons faites à partir des publications pour lesquelles le décalage dans le temps peut rendre certaines conclusions obsolètes.

4.3- Mesures de performance de classifieurs

4.3.1- Classification déterministe à deux classes

Nous considérons ici un problème simple de classification pour lequel nous nous intéressons à une classe unique C et nous voulons évaluer un système qui nous indique si un document peut être associé ou non à cette classe *C.* Ce problème est un problème de classification à deux classes (C et *non C* noté ¬ C). Si on peut maitriser ce problème simple, on pourra fusionner par la suite, les mesures de performance de plusieurs systèmes bi-classes afin d'obtenir une mesure de la performance d'un classifieur multi-classes.

4.3.1.1- Matrice de contingence

Pour évaluer un système de classification de ce type, nous utilisons un corpus étiqueté de documents (corpus d'apprentissage) pour lequel on connaît la vraie catégorie de chaque document, et le résultat obtenu par le classifieur. Pour ce corpus, nous pouvons construire la **matrice de contingence** pour chaque classe (Voir tableau 4.1), qui fournit 4 informations essentielles :

- Vrai Positif (**VP**) : Le nombre de documents attribués à une catégorie convenablement. (Documents attribués à leurs vraies catégories)
- Faux Positif (**FP**) : Le nombre de documents attribués à une catégorie inconvenablement. (Documents attribués à des mauvaises catégories)
- Faux Négatif (**FN**) : Le nombre de documents inconvenablement non attribués. (Qui auraient dû être attribués à une catégorie mais qui ne l'ont pas été).
- Vrai Négatif(**VN**) : Le nombre de documents non attribués à une catégorie convenablement (Qui n'ont pas à être attribués à une catégorie, et ne l'ont pas été)

Catégorie Ci		***Jugement Expert***	
		Oui	Non
Jugement classifieur	Oui	$\mathbf{VP_i}$	FP_i
	Non	FN_i	$\mathbf{VN_i}$

Tableau 4.1 : Matrice de contingence de la classe C_i

A partir de ce tableau de contingence, la communauté du TALN calcule divers indicateurs de base, eux-mêmes combinés pour donner d'autres mesures.

4.3.1.2- Précision et Rappel

Certains principes d'évaluation sont utilisés de manière courante dans le domaine de catégorisation de textes. Les performances en termes de classification sont généralement mesurées à partir de deux indicateurs traditionnellement utilisés c'est les mesures de rappel et précision. Initialement elles ont été conçues pour les systèmes de recherche d'information, mais par la suite la communauté de classification de textes les a adoptées.
Formellement, pour chaque classe C_i, on calcule deux probabilités qui peuvent être estimées à partir de la matrice de contingence correspondante, ainsi ces deux mesures peuvent être définies de la manière suivante :

- **Le rappel** étant la proportion de documents correctement classés dans par le système par rapport à tous les documents de la classe C_i.

$$Rappel\ (C_i) = \frac{Nombre\,de\,documents\,bien\,classés\,dans\,C_i}{Nombre\,de\,documents\,de\,la\,classe\,C_i}$$

$$R_i = \frac{VP_i}{VP_i + FN_i}$$

Le rappel mesure la capacité d'un système de classification à détecter les documents correctement classés. Cependant, un système de classification qui considérerait tous les documents comme pertinents obtiendrait un rappel de 100%. Un rappel fort ou faible n'est pas suffisant pour évaluer les performances d'un système. Pour cela, on définit la précision.

- **La précision est** la proportion de documents correctement classés parmi ceux classés par le système dans C_i .

$$Précision\ (C_i) = \frac{Nombre\,de\,documents\,bien\ classés\,dans\,C_i}{Nombre\,de\,documents\,classés\,dans\,C_i}$$

$$P_i = \frac{VP_i}{VP_i + FP_i}$$

La précision mesure la capacité d'un système de classification à ne pas classer un document dans une classe, un document qui ne l'est pas. Comme elle peut aussi être interprétée par la probabilité conditionnelle qu'un document choisi aléatoirement dans la classe soit bien classé par le classifieur.

Ces deux indicateurs pris l'un indépendamment de l'autre ne permettent d'évaluer qu'une facette du système de classification : la qualité ou la quantité. Les courbes de *précision* vs *rappel* permettent de mieux comprendre le comportement du classifieur, et de visualiser l'évolution de la précision en fonction du rappel pour les 11 niveaux standard [0-0,1-0,2-...-1].

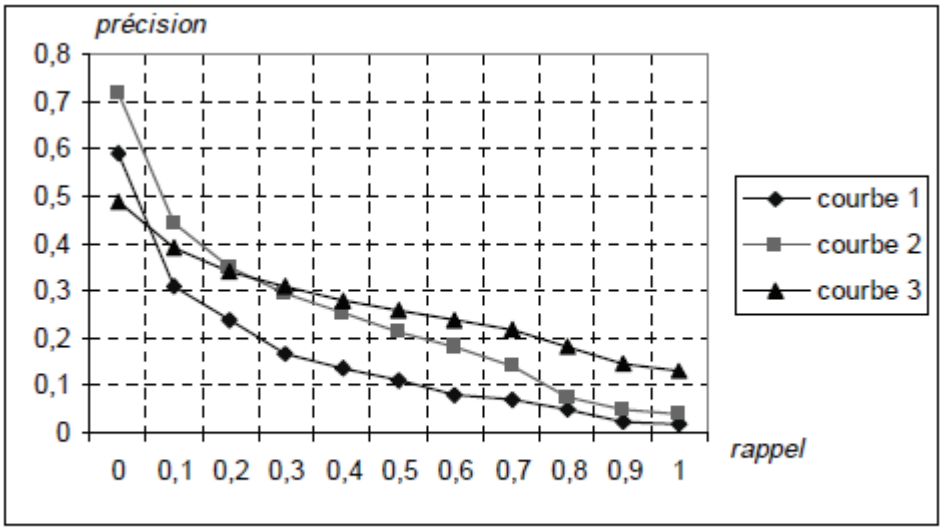

Figure 4.1 : Courbe Rappel-Précision pour trois classifieurs

Ces deux notions sont souvent utilisées dans le domaine de la recherche d'information, car elles reflètent le point de vue de l'utilisateur : si la précision est faible, l'utilisateur sera insatisfait, car il devra perdre du temps à lire des informations qui ne l'intéressent pas. Si le rappel est faible, l'utilisateur n'aura pas accès à une information qu'il souhaitait avoir.
Un classifieur parfait doit avoir une précision et un rappel de un (1), mais ces deux exigences sont souvent contradictoires et une très forte précision ne peut être obtenue qu'au prix d'un rappel faible et vice-versa.

4.3.1.3- Bruit et silence

On peut également définir les notions de *Bruit* (*B*) et de S*ilence (S)* qui sont respectivement les notions complémentaires de la précision et du rappel.
On utilise aussi la notion de bruit qui présente le problème selon le point de vue opposé de la précision. Le bruit est le pourcentage de textes incorrectement associés à une classe par le système :

$$Bruit\ (B)\ = 1 - Précision(P)\ = \frac{FP_i}{VP_i + FP_i}$$

La notion de silence est le point de vue opposé du rappel. Le silence est le pourcentage de textes à associer à une classe incorrectement non classés par le système :

$$Silence\ (S)\ = 1 - Rappel(R) = \frac{FN_i}{VP_i + FN_i}$$

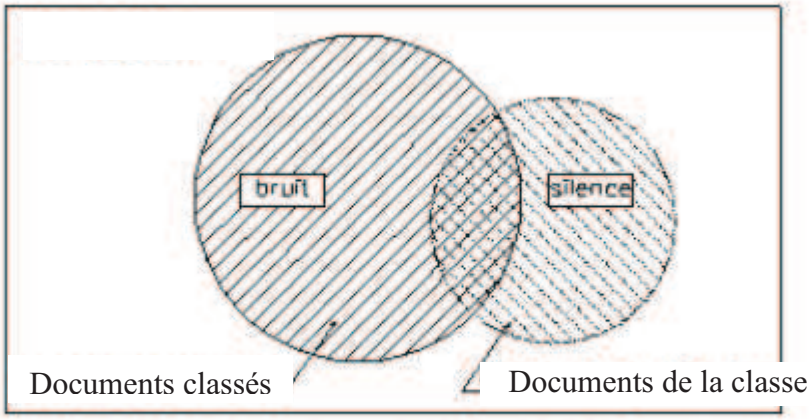

Figure 4.2 : Notions de bruit et de silence

4.3.1.4- Taux de succès et taux d'erreur

Le taux de succès ou l'exactitude Acc (Accuracy rate) et le taux d'erreur Err (Error rate) sont deux mesures souvent utilisées par la communauté de l'apprentissage automatique. Le taux de succès désigne le pourcentage d'exemples bien classés par le classifieur, tandis que le taux d'erreur désigne le pourcentage d'exemples mal classés.
Les deux taux sont estimés comme suit :

$$Acc = \frac{VP + VN}{VP + VN + FP + FN} \quad , \quad Err = \frac{FP + FN}{VP + VN + FP + FN} = 1 - Acc$$

4.3.1.5- Taux de chute et la spécificité

Deux autres indicateurs peuvent être utilisés pour mesurer la performance d'un classifieur :

$$Taux\ de\ chute = \frac{FP_i}{FP_i + VN_i}$$

$$Spécificité = \frac{VN_i}{FP_i + VN_i}$$

4.3.1.6- L'overlap et la généralité

$$Overlap = \frac{VP_i}{VP_i + FP_i + FN_i}$$

$$Généralité = \frac{VP}{VP + VN + FP + FN}$$

4.3.1.7- F-measure

Observés conjointement, les indicateurs les plus célèbres à savoir le rappel et la précision, sont une estimation courante de la performance d'un système de classification. Cependant plusieurs mesures ont été développées afin de synthétiser cette double information. Nous ne

retiendrons ici la mesure F_β décrite dans (Van Rijsbergen, 1979) . La ***F***-mesure est la mesure de synthèse communément adoptée depuis les années 80 pour évaluer les algorithmes de classification de données textuelles à partir de la précision et du rappel. Elle est employée indifféremment pour la classification (Non supervisé) ou la catégorisation (Supervisé), pour la problématique de recherche d'information ou de classification.
Elle permet donc, de combiner, selon un paramètre β, rappel et précision.
On définit la mesure F_β comme la moyenne harmonique entre le rappel et la précision :

$$F_\beta = \frac{(\beta^2 + 1) * précision * rappel}{\beta^2 * précision + rappel}$$

Pour utiliser cette mesure, il est donc nécessaire de fixer préalablement un seuil de décision pour le classement, et de calculer la valeur de F_β pour ce seuil.
Le paramètre β permet de choisir l'importance relative que l'on souhaite donner à chaque quantité. On choisit en général de donner la même importance aux deux critères, donc habituellement, la valeur de β est fixée à 1 et la mesure est ainsi notée F_1 (noté ***F)*** qui s'écrit :

$$F_1 = \frac{2 * P * R}{P + R}$$

Une des propriétés intéressante de cette mesure est le fait que, si $P = R = X$, alors $F = X$; cette mesure a alors une interprétation simple.

Afin de mieux comprendre le fonctionnement de ces 3 mesures (précision, rappel et mesure F), nous détaillons dans les tableaux 4.2 les performances de plusieurs classifieurs utilisés sur deux classes C et ¬ C de respectivement 100 et 200 documents.

Classifieur \ Expert	C	¬ C
C	100	200
¬ C	0	0
Rappel = 100 %		
Précision = 33 %		
F_1 = 50 %		

Tous les textes sont classés dans C

Classifieur \ Expert	C	¬ C
C	0	0
¬ C	100	200
Rappel = 0 %		
Précision = 100 %		
F_1 = 0 %		

Aucun texte n'est classé dans C

Classifieur \ Expert	C	¬ C
C	100	0
¬ C	0	200
Rappel = 100 %		
Précision = 100 %		
F_1 = 100 %		

Classifieur parfait

Classifieur \ Expert	C	¬ C
C	0	200
¬ C	100	0
Rappel = 0 %		
Précision = 0 %		
F_1 = 0 %		

Classifieur « le pire »

Tableaux 4.2 : Différents classifieurs et les mesures rappel, précision et F_1 associées

Les deux classifieurs du haut représentent des classifieurs « radicaux » qui classent tous les textes dans C, soit aucun n'est classé dans C. Les deux autres classifieurs sont des classifieurs, soit parfait – les textes sont correctement classés – soit dramatique – les textes sont toujours mal classés.

4.3.2- Classification déterministe à plusieurs classes

Pour la catégorisation à plusieurs classes de textes, une approche commune consiste à couper le processus de catégorisation en sous-problèmes. Chaque sous-problème concerne uniquement une classe et l'objectif est alors de juger si le nouveau texte appartient ou n'appartient pas à cette classe par opposition aux autres.

Pour la catégorisation multi-classes de textes, nous avons un ensemble de classes C = (C_1,..., $C_{/C/}$) où /C/ est le nombre de classes (|C| > 2). Nous notons N*i* le nombre de documents de C_i.
Pour chacune des classes, nous pouvons calculer comme précédemment le rappel, la précision et la mesure F_1, notés respectivement R_i , P_i et $F1_i$. Nous pouvons donc obtenir des mesures globales pour le système à |C| classes en moyennant ces mesures par classe.
La précision et le rappel globaux, c-à-d, sur toutes les classes peuvent être calculés à travers une moyenne des résultats obtenus pour chaque catégorie.

Cependant, si les classes ne possèdent pas le même nombre de documents, ces moyennes risquent de ne pas refléter la performance du classifieur pour les grandes classes. Les résultats de chaque catégorie peuvent être combinés de deux manières :

- On peut calculer un score pour chaque catégorie à partir de sa matrice de contingence puis déterminer la moyenne des scores sur l'ensemble des catégories **(macro-averaging)**. Dans ce cas, toutes les catégories interviennent de la même manière dans le calcul du score final quelque soit le nombre de documents qu'elles contiennent.

- Une autre possibilité est de créer une table de contingence globale pour toutes les catégories **(micro-averaging)** : le contenu d'une cellule de cette table correspond à la somme des valeurs de la même cellule dans la table de chaque catégorie (Yang, 1999).

4.3.2.1- Matrice de contingence globale

		Expert	
	$C = \{c_1, c_2, ..., c_{\|C\|}\}$	C_i	$\neg C_i$
Classifieur	C_i	$VP = \sum_{i=1}^{\|C\|} VP_i$	$FP = \sum_{i=1}^{\|C\|} FP_i$
	$\neg C_i$	$FN = \sum_{i=1}^{\|C\|} FN_i$	$VN = \sum_{i=1}^{\|C\|} VN_i$

Tableau 4.3 : Table de contingence globale

4.3.2.2- La micro-moyenne

Les mesures de type *micro moyenne* (ou *micro*) correspondent à une moyenne qui pondère chaque classe par son effectif.
La micro-moyenne (traduction de micro-averaging) calcule les mesures rappel et précision de façon globale : si l'on considère les tables de contingences associées à chaque catégorie, cela revient à sommer les cases VP, FP, FN et VN de chaque catégorie pour obtenir la table de contingence globale (voir le tableau 4.3).

Les différentes mesures sont ensuite calculées à partir des valeurs cumulées. La micro-moyenne accorde donc des poids importants aux catégories ayant beaucoup d'exemples. La performance du classifieur dépend surtout de sa capacité à traiter les catégories les plus fréquentes. Ainsi, la précision micro-moyenne et le rappel micro-moyenne sont estimés comme suit :

$$\boldsymbol{P} = \frac{VP}{VP + FP} = \frac{\sum_{i=1}^{|C|} VP_i}{\sum_{i=1}^{|C|} \left(VP_i + FP_i\right)}$$

$$\boldsymbol{R} = \frac{VP}{VP + FN} = \frac{\sum_{i=1}^{|C|} VP_i}{\sum_{i=1}^{|C|} \left(VP_i + FN_i\right)}$$

4.3.2.3- La macro-moyenne

Les mesures de type *macro moyenne* correspondent à une moyenne qui ne prend pas en compte la taille des classes.
La macro-moyenne (traduction de macro-averaging) évalue d'abord indépendamment chaque catégorie. Ensuite, la performance globale du classifieur est calculée en faisant la moyenne des mesures individuelles. Les différentes catégories ont alors la même importance. La précision et le rappel macro-moyenne sont calculés comme suit :

$$\boldsymbol{P} = \frac{\sum_{i=1}^{|C|} P_i}{|C|} \quad , \quad \boldsymbol{R} = \frac{\sum_{i=1}^{|C|} R_i}{|C|}$$

Ainsi, les mesures de type *micro moyenne* permettent d'obtenir une estimation du système en privilégiant les classes de grande taille tandis que les mesures de type *macro moyenne* donnent une information quant aux performances d'un système sur les petites classes.

Le tableau 4.4 résume les 6 mesures obtenues :

Mesure du rappel	*Rappel macro moyenne*	$\frac{\sum_{i=1}^{\|c\|} rappel_i}{\|C\|}$
	Rappel micro moyenne	$\frac{\sum_{i=1}^{\|C\|} n_i * rappel_i}{\sum_{i=1}^{\|C\|} n_i}$
Mesure de la précision	*Précision macro moyenne*	$\frac{\sum_{i=1}^{\|c\|} précision_i}{\|C\|}$
	Précision micro moyenne	$\frac{\sum_{i=1}^{\|C\|} n_i * précision_i}{\sum_{i=1}^{\|C\|} n_i}$
Mesure F_1	*F1 macro moyenne*	$\frac{\sum_{i=1}^{\|c\|} F_{1_i}}{\|C\|}$
	F1 micro moyenne	$\frac{\sum_{i=1}^{\|C\|} n_i * F_{1_i}}{\sum_{i=1}^{\|C\|} n_i}$

Tableau 4.4 : Les mesures de performances en classification multi-classes
Cette figure illustre les formules de calcul pour les mesures micro-moyenne et macro-moyenne

4.3.2.4- Une mesure issue de TREC : l'utilité

Les fonctions d'utilité ont été introduites dans le cadre de la tâche de filtrage, lors de la compétition TREC, décrite brièvement en annexe.
L'idée consiste à donner un nombre positif de points au système pour chaque document correctement classés et à retirer des points négatifs pour chaque document incorrectement classés. L'utilité est donc de la forme :

$$U = a.VP + b.FP$$

Où *VP* est le nombre de documents correctement classés, et *FP* est le nombre de documents incorrectement classés. Les coefficients *a* et *b* varient selon l'importance relative que l'on souhaite donner à chaque terme. Les valeurs les plus couramment utilisées sont a = 3, b = -2 et a = 3, b= -1.
L'évaluation de l'utilité ne nécessite que l'observation des documents classés ; elle est donc plus facilement calculable que le rappel.
Néanmoins, cette mesure présente quelques inconvénients qui ne sont pas détaillés dans notre mémoire, qui font qu'elle est peu utilisée en dehors de la conférence TREC (Voir Annexe).

4.3.3- Classification floue ou Ranking

Certains systèmes de classification, et notamment les classifieurs probabilistes ou ceux basés sur le calcul de distance, trient les catégories les plus adéquates dans l'ordre pour y classer le

texte. Les catégories sont classées soit par les distances croissantes ou par probabilités décroissantes.
Ils existent des mesures de performances adéquates à ces systèmes, inspirées de la recherche d'information adaptée à la classification; citons parmi lesquelles la technique du « 11-point average précision : **Précision moyenne sur 11 points** ».
Cette approche consiste à évaluer la précision et le rappel pour chacun des 11 seuils de 0 % à 100 % par pas de 10 % {0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}, puis de calculer les précisions et rappel moyens. La moyenne de ces 11 valeurs de précision estime la capacité de catégoriser un document. La moyenne de ces résultats obtenus pour les différents textes de l'ensemble de test permet d'évaluer la capacité globale du classifieur sur ce corpus.

4.4- Autres critères de comparaison de classifieurs

La majorité des travaux de développement concernant la classification et la recherche d'information s'appuient principalement, pour tirer des conclusions sur les classifieurs, sur des notions de rappel et de précision. Or, il existe d'autres critères pour évaluer et comparer deux systèmes. Si ces deux informations sont très utiles pour l'évaluation des performances des systèmes de classification, cela ne renseigne rien, par exemple sur la complexité ou la facilité d'utilisation du système, le temps de réponse, l'effort fourni par l'utilisateur, ou la présentation du résultat, ou encore d'autres facteurs d'évaluation des classifieurs qui sont introduits par la communauté d'Apprentissage Automatique indépendants de l'adéquation aux données d'apprentissage. Citons parmi eux :

- ***Compéhensibilité*** : Le modèle est-il compréhensible ? Le système donne-t-il des réponses permettant de comprendre pourquoi un document a été classé dans une certaine catégorie ou bien s'agit-il d'une fonction numérique calculée à partir de données servant d'exemples (Boite noire) ? La distinction principale entre induction (apprentissage) numérique et apprentissage symbolique inductif réside dans l'expression de la fonction f; l'apprentissage symbolique produit des expressions compréhensibles, telles que des règles de production ou des arbres de décision.
- ***Simplicité*** : apprécie le taux de simplicité des résultats d'apprentissage produits par le classifieur.
- ***Intelligibilié*** : évalue le degré d'intelligence du classifieur.
- ***Le temps de réponse et d'indexation*** : est aussi un point qui peut être fondamental.
- ***L'encombrement du système et les ressources en mémoire requises*** : l'espace alloué en mémoire vive et sur le disque dur qui doit être prise en compte dans de nombreux cas.

On peut trouver dans la littérature tous ces critères d'évaluation mais pas avec le même degré d'importance. Par exemple, (Dumais & all, 1998) ont mis en compétitions cinq techniques d'apprentissage selon trois critères :

- Training efficiency (Efficacité d'apprentissage) : Le temps moyen nécessaire pour l'apprentissage.
- Classification efficiency (Efficacité de classement) : Le temps moyen nécessaire pour la classification de nouveaux documents.
- Effectiveness (Capacité d'apprentissage) : L'aptitude de traitement des grands corpus d'apprentissage.

4.5- Conclusion

Nous avons montré dans ce chapitre que les mesures absolues de performances ont une portée limitée. Cette limitation est due, d'une part, à l'impossibilité de définir précisément la notion

de pertinence, et d'autre part, à l'impossibilité d'obtenir des corpus de grande taille totalement et correctement étiquetés.
Il est nécessaire de mesurer les performances d'un filtre sur un ensemble de thèmes pour d'une part limiter l'impact des erreurs d'annotations et d'autre part, pour juger globalement une approche sur des thèmes de difficultés différentes. Plusieurs indicateurs de mesures sont proposés dans la littérature et la recherche d'autres mesures plus fiables n'a pas cessé et reste toujours une matière intéressante pour les chercheurs.
Néanmoins toutes les mesures présentées dans ce chapitre traitent toutes les erreurs avec la même importance, alors que du point de vue de l'utilisateur, cette assertion n'est pas vraie.

Il est cependant très difficile de prendre ces informations en considération, puisqu'il existe une grande part de subjectivité dans ces appréciations, et que finalement la seule vraie mesure est la satisfaction de l'utilisateur.