Chapitre 4 Évaluation des performances d'un filtre

Pour comparer les différents systèmes de filtrage, il faut définir une mesure pour évaluer leurs performances. Malheureusement, différentes mesures sont utilisées dans la littérature, ce qui rend les comparaisons souvent difficiles. Le but de ce chapitre n'est pas de les présenter toutes, mais d'introduire uniquement les plus importantes d'entre elles, et de mettre en évidence la difficulté de l'évaluation des performances.

4.1 Mesures de la précision et du rappel

4.1.1 Définitions

La précision et le rappel sont deux quantités qui sont définies lorsque les filtres prennent des décisions binaires : soit un document est sélectionné par le filtre, soit il ne l'est pas. Lorsque les ensembles de documents pertinents et non pertinents sont connus sur un corpus, il est alors possible d'évaluer les quantités définies à la Figure 4.1.

	Pertinents	Non pertinents
Sélectionnés	*a*	*b*
Non sélectionnés	*c*	*d*
Total	*P*	*NP*

Figure 4.1 : *Table de contingences pour un filtre binaire.*

Le rappel R et la précision P et sont définis par :

$$R = \frac{a}{a + c} \qquad P = \frac{a}{a + b}$$

Le *rappel* est le rapport du nombre de documents pertinents trouvés par le filtre au nombre de documents pertinents disponibles. Il s'agit de la proportion de documents bien classés pour la classe des documents pertinents : c'est une mesure utilisée habituellement en classification.

La *précision* est la proportion de documents pertinents parmi les documents sélectionnés. Cette quantité ne représente pas un taux d'exemples bien classés par rapport à une classe et n'est donc pas normalisée.

Ces deux notions sont souvent utilisées, 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 filtre parfait doit avoir une précision et un rappel de un, 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.

En effet, dans le cas limite où aucun document n'est sélectionné, la précision vaut un et le rappel est nul. Dans le cas limite où tous les documents sont sélectionnés, le rappel vaut un et la précision est de $\frac{P}{P + NP}$; cette quantité, appelée *densité* du thème, est généralement assez faible, et représente la précision moyenne que l'on obtiendrait en sélectionnant les documents aléatoirement.

On peut également définir les notions de *bruit* (*B*) et de *silence (S)* qui sont respectivement les notions complémentaires de la précision et du rappel :

$$B = 1 - P \qquad S = 1 - R$$

4.1.2 Une estimation pas si simple

Dans la pratique, les valeurs exactes de la précision et du rappel ne peuvent pas être calculées et, par conséquent, les valeurs absolues n'ont pas de sens.

Pour mesurer les performances d'un filtre (comme celles de tout classifieur), il faut utiliser une base de test indépendante de la base d'apprentissage. Pour que les résultats obtenus sur cette base de test aient un sens, deux conditions doivent être remplies : il faut, d'une part, que cette base soit suffisamment grande et représentative, et, d'autre part, il faut pouvoir évaluer les quantités *a*, *b*, *c* et *d* sur cette base. Or la connaissance de ces quantités nécessite la catégorisation manuelle de chaque document de la base, ce qui est justement très difficile à faire si la base est grande.

Ainsi, sur la base de test utilisée pour TREC-8 qui comporte 140.000 documents, seule une partie de ces documents a été examinée par des assesseurs. Tous ceux qui n'ont pas été examinés sont considérés comme non pertinents (la manière dont les documents ont été étiquetés a été présentée au chapitre précédent et est détaillée dans [Voorhees et Harman, 2000]). Il peut donc exister des documents pertinents sur la partie de documents non examinée.

Sur le corpus Reuters, l'ensemble des documents de la base de test est supposé avoir été correctement affecté "manuellement". En fait, dans la pratique il n'en est rien : certains documents pertinents sont étiquetés comme non pertinents. Par exemple, la Figure 4.2 montre un texte étiqueté comme non pertinent alors qu'il est pertinent pour la catégorie *money, foreign-exchange*, comme le montre le passage qui figure en gras.

MIYAZAWA SEES EVENTUAL LOWER U.S. TRADE DEFICIT
Japanese Finance Minister Kiichi Miyazawa told a press conference he expects the U.S. Trade deficit to eventually start reflecting economic fundamentals, which should influence exchange rates.
The minister was not referring to the U.S. Trade data to be released in Washington later today.
Miyazawa also said he told major industrial nations when he was in Washington last week that **present exchange rates are not necessarily good. He had said earlier in Washington that current exchange rates were within levels implied in the February Paris currency accord**.
REUTER

Figure 4.2 : *Texte 16745 du corpus Reuters. Ce texte est pertinent pour la catégorie money, foreign-exchange alors qu'il a été étiqueté manuellement comme non pertinent.*

De même le texte 21477 de la base de test présenté à la Figure 4.3 n'est pas étiqueté comme pertinent pour le thème *interest* alors que le passage en gras prouve qu'il est, en fait, pertinent.

DEUTSCHE BANK CHIEF SAYS LOUVRE PACT STILL INTACT
Deutsche Bank AG joint chief executive Friedrich Wilhelm Christians said he believed the Louvre accord on currency stability was still intact.
Christians told a news conference he met U.S. Treasury Secretary James Baker in the last two weeks, after **short term German interest rates had risen twice**.
"I am sure that with 1.7720 marks the dollar is still within the Louvre agreement. I do not see that the accord has been terminated," Christians said. He was responding to questions about comments by Baker, who said the Louvre accord was still operative but criticised rises in West German interest rates.
REUTER

Figure 4.3 : *Texte 21477 du corpus Reuters. Ce texte est pertinent pour la catégorie money, interest alors qu'il a été étiqueté manuellement comme non pertinent.*

La pertinence ou non d'un document dépend également de la personne qui étiquette les documents : d'une personne à l'autre, le même document peut être déclaré comme pertinent ou non pertinent. Ceci est surtout vrai pour la catégorisation de documents, puisqu'il n'existe pas, en général, de définition très précise du thème. Ainsi le texte présenté en Figure 4.4 peut être considéré comme pertinent pour le thème des *participations* ou non pertinent selon la définition que chacun donne exactement à ce thème.

```
Les banques italiennes san Paolo di Torino et Istituto Mobiliare Italiano
(IMI) ont signé lundi à Turin l'acte formel de fusion de leurs deux
établissements, qui donnera naissance à la première banque italienne.
```

Figure 4.4 : *Exemple de texte difficile à classer pour le thème participation.*

Pour toutes les raisons décrites ci-dessus, les valeurs exactes de précision et de rappel ne sont pas accessibles, et, en tout état de cause, leurs valeurs peuvent varier selon les personnes qui jugent les documents. En pratique les valeurs absolues n'ont donc pas beaucoup de sens ; en revanche, les valeurs relatives ont un sens pour comparer des systèmes entre eux, puisque les évaluations sont faites avec les mêmes approximations.

Cependant, pour que les approximations n'avantagent pas artificiellement un système par rapport à l'autre, il est indispensable de moyenner les performances de chaque système sur un ensemble de thèmes différents. On verra dans le paragraphe 4.4 comment agréger les résultats pour un ensemble de thèmes.

4.2 Courbes rappel-précision

4.2.1 Courbes non interpolées

En général, les filtres statistiques fournissent une probabilité de pertinence pour chaque document ; pour en déduire une réponse binaire, il faut déterminer une valeur pour le seuil de décision utilisé. Il est donc possible de calculer la précision et le rappel correspondant à chaque valeur du seuil de décision et de tracer l'évolution de ces deux quantités.

4.2.2 Courbes interpolées

On préfère calculer la précision pour des valeurs prédéfinies du rappel, de 0 % à 100 % par pas de 10 %. En pratique ces valeurs du rappel peuvent ne pas être atteintes exactement : les valeurs de la précision doivent donc être interpolées. La règle d'interpolation est la suivante : la

valeur interpolée de la précision pour un niveau de rappel *i* est la précision maximale obtenue pour un rappel supérieur ou égal à *i.* Cette règle d'interpolation définit donc également une précision pour un rappel nul alors qu'une telle valeur n'existe pas. La Figure 4.5 montre la courbe obtenue pour le filtre précédent avec les valeurs interpolées.

Cette courbe montre qu'il est toujours possible d'obtenir une précision élevée au prix d'un rappel faible ou un rappel élevé au prix d'une précision faible. Dans la pratique, on essaye de choisir un compromis entre ces deux exigences. On verra néanmoins dans le chapitre 9 que, pour certaines applications, le filtre fonctionne avec une précision élevée au détriment du rappel.

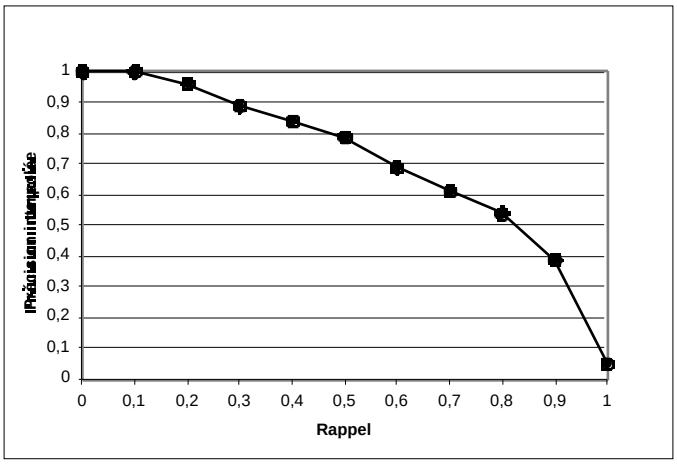

Figure 4.5 : *Courbe rappel-précision interpolée.*

L'avantage de cette interpolation est qu'elle permet de connaître la précision pour des valeurs standardisées. Lorsque plusieurs thèmes sont étudiés, on peut facilement obtenir la courbe moyenne d'un système en moyennant simplement toutes les précisions obtenues aux différents seuils du rappel pour les différents thèmes. Les performances d'un filtre sur un ensemble de thèmes peuvent donc être caractérisées par une seule courbe.

4.3 Caractérisation d'un système binaire par une seule grandeur

Le paragraphe précédent a montré qu'un système est caractérisé par une courbe ou par un couple (rappel, précision). Dans la pratique, il n'est pas facile de comparer les filtres sur la

base de ces caractéristiques ; on cherche donc à évaluer leurs performances par un seul nombre. Nous présentons trois mesures parmi les plus utilisées, en précisant leurs inconvénients éventuels. Il ne faut pas perdre de vue que la caractérisation d'une courbe par un seul nombre implique nécessairement une simplification ; lorsque cela est possible, il peut être avantageux d'utiliser tout de même les courbes rappel-précision

4.3.1 Précision moyenne sur 11 points

La précision moyenne sur 11 points consiste simplement à moyenner les 11 précisions interpolées obtenues pour les seuils de rappels fixes définis, de 0 %, à 100 % par pas de 10 %.

4.3.2 Point moyen

Le point moyen est défini comme le point pour lequel la précision est égale au rappel : c'est l'intersection de la courbe rappel-précision avec la diagonale principale.

S'il n'existe pas de seuil permettant de satisfaire cette égalité, une extrapolation linéaire est effectuée : si (R_1, P_1) et (R_2, P_2) sont les deux couples qui encadrent le point moyen, une extrapolation linéaire permet de trouver le point moyen :

$$R = P = \frac{R_2 P_1 - R_1 P_2}{R_2 - R_1 + P_1 - P_2}$$

Malheureusement, la valeur obtenue grâce à cette extrapolation n'a pas de signification réelle : elle caractérise plus une propriété d'un point de la courbe rappel-précision que la qualité qu'un système.

Nous éviterons cette mesure dans la suite de ce mémoire.

4.3.3 Mesure *F*

La mesure F_β [van Rijsbergen, 1979] prend en considération la précision et le rappel simultanément. Elle est définie par :

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 : on utilise F_1 (noté F dans toute la suite de ce mémoire) qui s'écrit :

$$F = \frac{2.P.R}{P + R}$$

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.

Lorsque l'on utilise la mesure F, il ne faut pas perdre de vue qu'une partie de l'information est perdue, puisque cette mesure rend compte d'un seul point de la courbe dans la zone particulière où la précision et le rappel sont du même ordre de grandeur. La Figure 4.6 montre les résultats obtenus pour le thème *interest* du corpus Reuters pour trois jeux de paramètres différents. Avec la mesure F, ces trois systèmes ont des performances similaires, mais les courbes font apparaître des différences qui ne sont pas mises en évidence lorsque l'on caractérise les courbes par un seul point. D'après ces courbes, le système ayant la mesure de F la plus faible (F = 65,5), semble présenter le meilleur compromis : en effet, pour des valeurs de rappel faible, il obtient une précision élevée aussi élevée que le troisième système, et, dans la zone de rappel élevée, il a une précision comparable à celle du premier système.

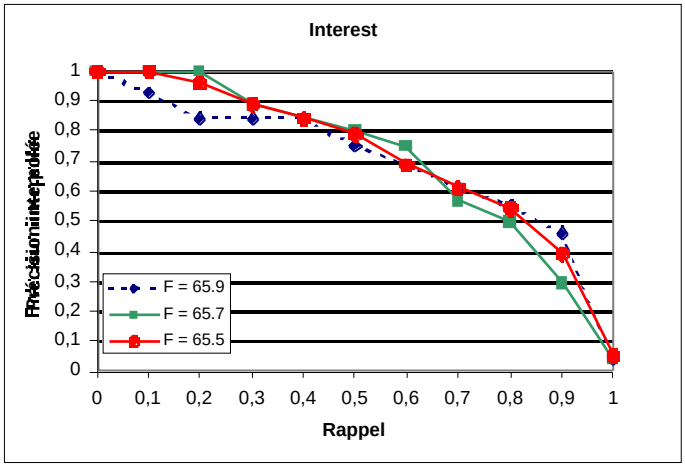

Figure 4.6 : *Courbes rappel-précision interpolée pour trois systèmes différents pour le thème interest du corpus Reuters.*

Malgré ces imprécisions, cette mesure présente un grand intérêt pratique, et elle est souvent utilisée dans les publications ; nous l'utiliserons fréquemment dans la suite de ce mémoire.

En pratique, la mesure de *F* est une fonction du seuil de décision du filtre ; par conséquent, afin de s'affranchir du choix du seuil de décision, il est possible de tester plusieurs seuils de 0 à 1 par pas de 0,1, et de conserver le seuil correspondant à la valeur de *F* la plus élevée sur la base de test. Cette mesure est appelée *F optimal* et représente la borne supérieure de la valeur de *F* que le système peut atteindre dans la pratique.

4.4 Moyenne sur un ensemble de thèmes : macro-moyenne et micro-moyenne

Comme les différences entre les performances de deux systèmes peuvent être liées aux incertitudes sur les mesures, il est indispensable d'effectuer une comparaison sur un ensemble assez vaste de thèmes, et de calculer une moyenne des performances sur cet ensemble de thèmes. Cette moyenne peut être calculée de deux manières.

- la macro-moyenne consiste à faire simplement une moyenne sur l'ensemble des thèmes des performances individuelles de chaque thème pour la mesure choisie. Cette moyenne donne un poids égal à chaque thème ;
- pour calculer la micro-moyenne, les tables de contingences de chaque thème, comme celle de la Figure 4.1, sont additionnées, puis les valeurs de précision et de rappel sont calculées sur cet ensemble, qui n'est plus considéré que comme un seul thème.

La différence de comportement des deux calculs se comprend bien sur un corpus ayant les caractéristiques du corpus Reuters.

Puisque toutes les catégories ont la même importance dans le calcul de la macro-moyenne, le résultat est surtout gouverné, sur ce corpus, par les catégories ayant peu de documents pertinents puisque celles-ci sont majoritaires.

En revanche, la micro-moyenne est gouvernée par les premières catégories (celles qui ont le plus de documents pertinents). Plus précisément, si un classifieur sélectionne tous les documents pertinents pour les vingt premières catégories et aucun document pour les catégories restantes, le rappel calculé pour la micro moyenne est de 83,9 %. Pour les trente premières catégories, le rappel est de 89,2%. Pour obtenir de bonnes performances avec la micro-moyenne, une bonne stratégie consiste donc à ne rien faire pour les soixante dernières

catégories puisque la perte de rappel est très faible et que le risque de faire chuter la précision est supprimé.

4.5 Précision moyenne non interpolée

Les mesures précédentes nécessitent l'utilisation d'un seuil de décision, alors que la précision moyenne non interpolée caractérise la qualité d'un classement. Le système calcule une probabilité de pertinence pour l'ensemble des documents qui constituent la base de test, et les classe par ordre de pertinence décroissante à la manière des moteurs de recherche sur le web.
En parcourant cette liste, la précision est calculée pour chaque document pertinent ; la précision moyenne non interpolée est obtenue en additionnant ces différentes précisions, et en divisant la somme par le nombre total de documents pertinents présents sur la base de test.

Pour fixer les idées, nous présentons un exemple. Supposons qu'une base de test comporte sept documents parmi lesquels trois sont pertinents, et qu'un filtre a permis d'obtenir le classement des documents présenté sur la Figure 4.7.

Classement	Pertinent	Précision
t1	non	-
t2	oui	1/2
t3	non	-
t4	oui	2/4
t5	non	-
t6	non	-
t7	oui	3/7

Figure 4.7 : Exemple de classement avec les précisions calculées pour chaque document pertinent.

Pour cet exemple, la **précision moyenne non interpolée** est (on utilise l'acronyme *UAP* pour désigner cette mesure, du nom anglais : *Uninterpolated Average Precision*) :

$$UAP = \frac{1}{3}(1/2 + 2/4 + 3/7) = 0,47$$

Ce nombre mesure la qualité du classement, grâce à un score évoluant entre 0 et 1, indépendamment du choix d'un seuil de décision.

4.6 Une mesure issue de TREC : l'utilité

4.6.1 Motivation et définition de l'utilité

Tous les critères d'évaluation présentés jusqu'ici présentent un inconvénient majeur : pour les calculer, il est nécessaire de connaître l'ensemble des documents pertinents. Or, dans la pratique, pour un système qui filtre les dépêches quotidiennement, il est impossible de calculer le rappel, puisque l'utilisateur ne dispose pas des dépêches pertinentes non sélectionnées.

De plus, considérons deux filtres différents pour un même thème, tels que le premier sélectionne 100 documents non pertinents et aucun document pertinent, alors que le deuxième sélectionne un document non pertinent et aucun document pertinent. Ces deux systèmes ont un rappel et une précision nuls ; néanmoins, dans la pratique, le deuxième filtre est préférable au premier puisqu'un utilisateur ne perd pas de temps à lire des documents qui ne l'intéressent pas. Dans ce cas, les mesures de précision et de rappel (et les mesures dérivées) ne permettent pas de différencier les deux systèmes.

Pour remédier à ces inconvénients, les fonctions d'utilité ont été introduites lors de la compétition TREC [Hull, 1999] dans le cadre de la tâche de filtrage.

Pour différencier les deux systèmes précédents, l'idée consiste à donner un nombre positif de points au système pour chaque document pertinent sélectionné et à retirer des points négatifs pour chaque document non pertinent sélectionné. L'utilité est donc de la forme :

$$U = a.P + b.NP$$

où P est le nombre de documents pertinents sélectionnés, et NP est le nombre de documents non pertinents sélectionnés. 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 sélectionnés ; elle est donc plus facilement calculable que le rappel.

4.6.2 Inconvénients de l'utilité

Cette mesure présente quelques inconvénients, qui font qu'elle est peu utilisée en dehors de la conférence TREC.

· Elle n'est pas facilement interprétable, contrairement à la précision et au rappel. Plus précisément, le lien avec ces deux notions n'est pas immédiat. En fait si l'on considère deux systèmes X et Y, il est même possible d'obtenir les caractéristiques suivantes :

$$P(X) > P(Y)$$

$$R(X) > R(Y)$$

et néanmoins :

$$U(X) < U(Y)$$

La démonstration peut être trouvée dans [Hull et Robertson, 2000].

· La définition de l'utilité n'est pas normalisée d'un thème à l'autre : l'utilité maximale pour un thème donné dépend du nombre de documents pertinents présents dans l'ensemble du corpus P_c puisque cette utilité maximale est $a.P_c$. Il est donc impossible de moyenner les résultats obtenus à travers différents thèmes. Une normalisation a été proposée [Hull, 1999], mais son utilisation n'est pas très simple.

- Pour la tâche du filtrage adaptatif de la compétition TREC-8, tous les systèmes avaient une utilité négative. En conséquence, la meilleure des stratégies consistait à utiliser un système qui ne faisait rien puisqu'un tel système aurait eu une utilité nulle.

Ces différentes considérations font que l'utilité est rarement employée en dehors de TREC. Dans l'avenir, cette mesure devrait évoluer pour pallier ces inconvénients.

4.7 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.

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 :

- lorsqu'un document non pertinent est sélectionné, l'utilisateur sera plus indulgent si le document est "proche" du thème que si le document n'a absolument rien à voir avec celui-ci ;
- il n'est pas très grave de ne pas sélectionner certains documents pertinents si l'information qu'ils contiennent a déjà été apportée par d'autres documents ; en revanche, si l'utilisateur n'est pas du tout informé d'une nouvelle qu'il apprend par ailleurs, sa confiance dans le système diminuera fortement.

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.

Dans la suite de ce mémoire, nous utiliserons principalement :

- les courbes rappel-précision, car elles apportent une information complète sur le comportement du filtre.
- la mesure *F*, car elle rend compte du comportement d'un filtre, une fois qu'un seuil de décision a été choisi : elle traduit la "perception" d'un utilisateur dans le cadre d'une application de filtrage.
- la précision moyenne non interpolée (notée *UAP*), car elle rend compte de la qualité du classement proposé par un filtre indépendamment du seuil de décision : un filtre peut avoir une valeur de *F* faible uniquement parce que le seuil de décision est mal choisi, et non parce que le classifieur est de mauvaise qualité.