FILTRAGE AUTO-ADAPTATIF BASE SUR L'ANALYSE DE LA VARIANCE.

KAROUACH Saïd,
karouach@irit.fr

DOUSSET Bernard,
dousset@irit.fr

BOUTILLAT Nicolas
boutillat@irit.fr

Adresse professionnelle
IRIT-SIG, Université Paul Sabatier, 118, route de Narbonne
31062 Toulouse cedex 04
Tél : 05.61.55.67.81

Résumé : Nous abordons, ici, le problème délicat du filtrage de l'information et de ses différentes applications en amont du processus de veille scientifique et technique. Toute recherche d'information nécessite une phase de ciblage et de validation qui permet un recentrage sur le sujet choisi. Notre laboratoire s'est depuis longtemps penché sur ce problème et a proposé des techniques de filtrage basées essentiellement sur les réseaux de neuronaux **[BOUG01] [TMAR01]** et la notion de profil. Nous reprenons cette approche afin de proposer une alternative basée sur l'analyse de la variance. Son principe est le suivant : en partant d'une liste de documents corrélés positivement ou négativement avec un thème donné (profil), nous proposons de filtrer de nouveaux documents, pris à la volée ou dans un corpus existant. Un jugement sur la pertinence des réponses permet de recaler le modèle proposé par analyse de la variance et donc d'affiner ou d'adapter le filtrage soit en continu (dépêches d'agence, push) soit en boucle (validation d'un corpus sur des extraits). Nous proposons deux approches : modèle tout ou rien et arbitrage par pondérations. La pertinence de cette approche sera illustrée par des exemples et des comparaisons à des techniques déjà existantes.

Abstract : In this paper, we discuss the information filtering problem and its various applications connected with science and technology watch. Any information retrieval requires a phase of targetting and validation which allows a centring on the selected subject. Our laboratory was for a long time concentrated on this problem and proposed techniques of filtering based primarily on the neuronal networks **[BOUG01] [TMAR01]** and the concept of profile. We take again this approach in order to propose an alternative based on the analysis of the variance. Its principle is as follows: on the basis of a list of documents correlated positively or negatively with a given topic (profile), we propose to filter new documents, taken with stolen or in an existing corpus. A judgement on the relevance of the answers makes

it possible to readjust the model suggested by analysis of the variance and thus to refine or adapt filtering either uninterrupted (dispatches of agency, push) or buckles some (validation of a corpus on extracts). We propose two approaches: all or nothing model and arbitration by weightings. The relevance of this approach will be illustrated for examples and comparisons with already existing techniques

Mots-clés : Filtrage - Data mining - Veille stratégique scientifique et technique - Analyse de données.

Keywords : Filtering - Data mining - Science and technology watch - Data analysis.

Filtrage auto-adaptatif basé sur l'analyse de la variance.

1 - LE PROBLEME POSE

Initialement, nous partons de deux collections de documents :

- Les documents de référence (le profil)
- Les documents à filtrer (flot de documents, corpus, …)

La première collection est évaluée par l'utilisateur

- En mode tout ou rien : un document est pertinent ou ne l'est pas
- En mode pondéré : un document est plus ou moins pertinent sur une échelle ayant au moins trois niveaux fixée a priori.

Le profil est analysé pour en extraire la terminologie significative. Chaque terme possède donc une fréquence dans le profil ainsi que dans chaque document du profil. Nous allons désigner par F_j la fréquence relative du terme j dans le profil et par f_{ij} la fréquence relative de ce même terme dans le document i. Nous définissons ensuite une fonction de validité v_j qui est linéaire par rapport aux fréquences relatives et dont la valeur doit nous permettre de nous prononcer sur la pertinence de chaque document. Pour les documents du profil, cette fonction vaut théoriquement :

- -1 pour les documents non pertinents et +1 pour les documents pertinents (en mode tout ou rien)
- le niveau de pertinence attribué par l'utilisateur (en mode pondéré)

Il nous reste à trouver le modèle linéaire ou affine qui vérifie au mieux l'ensemble des équations ainsi produites :

$$v_i = \beta_0 + \sum_{j=1}^{n} \beta_j f_{ij} + e_i$$

où m est le nombre d'équations (i=1,m et m>n) et où e_i représente l'erreur sur l'équation i, β_j les coefficients du modèle, β_0 le terme constant dans le cas d'un modèle affine.

2 -LES DEUX APPROCHES POSSIBLES

2.1 - Méthode des moindres carrés

La méthode des moindres carrés consiste à minimiser la somme des carrés des écarts constatés entre les résultats de mesure et les valeurs obtenues à l'aide du modèle linéaire ou affine.

$$S = \sum_{i=1}^{m} e_i^2 = \sum_{i=1}^{m} \left(v_i - \beta_0 - \sum_{j=1}^{n} \beta_j f_{ij}\right)^2$$

On peut aussi minimiser la racine carrée de S, qui n'est autre que la norme euclidienne du vecteur des erreurs dans R^m. Il s'agit en fait de trouver la pseudo solution du système linéaire à m lignes et n colonnes admettant comme second membre le vecteur des mesures de validité des documents du profil.

$$[f_{ij}][\beta_j] \cong [v_i] \quad (1)$$

La meilleure solution au sens des moindres carrés de ce système est obtenue en résolvant le système de Cramer associé :

$$[f_{ji}][f_{ij}][\beta_j] = [f_{ji}][v_i] \quad (2)$$

Mais dans notre cas, l'ensemble des fréquences relatives est lié par une équation et ce pour chacun des documents du profil :

$$\sum_{j=1}^{n} f_{ij} = 1$$

Cette particularité rend instable la résolution du système linéaire associé (2), nous avons donc opté pour la suppression de la dernière colonne ($j=n$) car la valeur f_{in} peut être déduite des précédentes par complément à 1.

2.2 - Méthode du maximum de vraisemblance

Comme précédemment, nous réalisons n mesures de fréquences relatives sur chacun des m documents du profil et nous devons obtenir les m résultats escomptés suivants:

$$v_i \quad i=1,m$$

Nous représenterons cette série d'évaluations par un point V de l'espace R^m.

$$V=(v_1,v_2,\cdots,v_m)$$

Nous avons donc dans R^m une fonction de répartition:

$$d=f(V)=f(v_1,v_2,\cdots,v_m)$$

De plus, cette fonction de répartition dépend des $n+1$ paramètres λ_j du modèle affine optimal que nous essayons de déterminer (un pour chaque fréquence relative calculée, un autre pour le terme constant) :

$$d=d(v_1,v_2,\cdots,v_m;\lambda_0,\lambda_1,\cdots,\lambda_n)$$

La probabilité de réalisation de V est:

$$d(v_1,v_2,\cdots,v_m;\lambda_0,\lambda_1,\cdots,\lambda_n)dv=d(v_1,v_2,\cdots,v_m;\lambda_0,\lambda_1,\cdots,\lambda_n)dv_1\,dv_2\cdots dv_m$$

Nous cherchons la probabilité de réalisation des m expériences en sachant qu'elles sont indépendantes et identiquement distribuées:

$$Prob(v_1\cap v_2\cap\cdots\cap v_m)=\prod_{i=1}^{m}Prob(v_i)$$

$$Prob(v_i)=f(v_i,\lambda_0,\lambda_1,\cdots,\lambda_n)dv_i$$

$$\prod_{i=1}^{m}Prob\ (v_i)=\prod_{i=1}^{m}f(v_i,\lambda_0,\lambda_1,\cdots,\lambda_n)\prod_{i=1}^{m}dv_i$$

La méthode du maximum de vraisemblance consiste à maximiser la probabilité de réalisation de la série d'expériences V. Pour cela nous pouvons maximiser le logarithme de cette probabilité afin de transformer les produits en sommes.

<u>Remarque</u> : cadre des hypothèses de l'analyse de la variance:

Nous allons supposer que les erreurs commises sur les expériences par le modèle affine vérifient les propriétés suivantes:

1. $Esp(e_i)=0$,
2. Les e_i sont des variables aléatoires indépendantes,
3. les e_i ont même variance σ^2.

Nous pouvons donc en déduire que les e_i ont un comportement de loi normale centrée de variance σ^2: $N(0,\sigma^2)$.

Si l'approximation par le modèle linéaire s'écrit:

$$v_i=\lambda_0+\sum_{j=1}^{n}\lambda_j f_{ij}+e_i\ \forall i=1,m$$

On en déduit que le v_i suivent aussi une loi normale:

$$v_i \approx N(\lambda_0 + \sum_{j=1}^{n} \lambda_j f_{ij}, \sigma_2)$$

La fonction de répartition des v_i peut alors s'écrire sous la forme:

$$f(v_i, \lambda_1, \lambda_2, \cdots, \lambda_n) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(v_i - \lambda_0 - \sum_{j=1}^{n} \lambda_j f_{ij})^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{e_i^2}{2\sigma^2}}$$

$$L = \frac{1}{(\sigma\sqrt{2\pi})^m} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{m} e_i^2}$$

$$S = Log(L) = -m Log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma_2}\sum_{i=1}^{m} e_i^2$$

Le problème consiste alors à trouver le maximum de *S*, ce qui revient à trouver le minimum :

$$Min(\sum_{i=1}^{m} e_i^2)$$

car:

$$-m\, Log(\sigma\sqrt{2\pi}) \quad \text{est une constante,} \quad \sum_{i=1}^{m} e_i^2 \geq 0 \quad \text{et} \quad \frac{1}{2\sigma^2} \geq 0$$

Nous sommes donc conduits à résoudre le même problème qu'avec la méthode des moindres carrés. Les estimateurs obtenus par ces deux méthodes sont donc rigoureusement identiques.

3 - PRINCIPE GENERAL DE L'OUTIL DE FILTRAGE

3.1 - Constitution du profil initial

Le principal problème d'un procédure de filtrage se situe lors de la phase de démarrage. En effet, il est difficile de définir à priori quel seront les représentants les plus pertinents (documents ou mots-clés) qui seront capables de bien cibler le sujet et surtout quels seront les éléments non pertinents qui permettront d'éliminer une part importante du bruit (faux amis, sujets voisins, domaines d'application indésirables, …). Une collection hétérogène initiale reste donc à trouver pour permettre le démarrage du processus, mais les premiers filtrages sont souvent catastrophiques, car toutes les sources d'erreurs n'ont pas été prévues. Le mieux est d'analyser une collection assez vaste et de sélectionner des documents qui ne sont pas voisins, afin de recouvrir aux mieux l'espace des possibilités. Cette phase peut parfaitement être réalisée depuis notre plate-forme Tétralogie. Mais, comme dans le cas d'un estimateur, le nombre de documents doit être supérieur au nombre de termes utilisés dans le modèle, nous n'hésitons pas à prendre une collection assez vaste (une bonne centaine de documents semble un point de départ honorable).

3.2 - Choix des termes discriminants

Dans notre cas, nous devons trouver un équilibre entre les terminologies positives (corrélées aux document acceptés) et négatives (corrélées aux documents rejetés). Une zone d'incertitude doit aussi être aménagée afin de conserver certains termes présents dans les deux groupes : on évite ainsi une source d'instabilité de la méthode. Pour effectuer ce choix nous pouvons utiliser notre plate-forme Tétralogie. Une première étape consiste à

extraire toute la terminologie du profil en évitant les mots vides. Ensuite chaque document est analysé pour déterminer la fréquence absolue de tous les termes retenus. Grâce au tableur on peut extraire la matrice des fréquences relatives et proposer une liste de termes caractéristiques de chaque groupe (acceptés ou rejetés) et de termes caractéristiques du profil sans discernement. La concaténation des trois listes est dédoublonnée et tronquée afin que le nombre de termes retenus soit inférieur au nombre de documents du profil. La meilleure solution est de prendre les termes les plus fréquents en évitant les termes équirépartis sur l'ensemble des documents.

3.3 - Filtrage des nouveaux documents

Une fois l'estimateur calculé (fonction affine des fréquences relatives), il suffit de calculer ces fréquences pour chaque document à filtrer et d'appliquer la fonction au vecteur associé. La valeur de la fonction est réelle mais elle s'approche soit du critère d'acceptation soit du critère de rejet.

Dans le cas du tout ou rien, les valeurs cibles sont +1 et –1, l'incertitude est donc au zéro. Suivant le nombre de documents acceptés on peut décaler la frontière en fonction de la capacité de lecture de l'utilisateur. Pour une utilisation électronique des documents filtrés, le problème ne se pose pas, il convient donc de tout garder au dessus de zéro.

Dans le cas d'une échelle de niveaux de pertinence, la restitution peut être plus nuancée. Les documents peuvent être triés ou classés en fonction de leur pertinence et toujours des capacités ultérieures de traitement. Il convient, dans ce cas, de conserver certains documents peu pertinents afin d'alimenter le profil pour bien tenir compte de la répartition en niveaux.

3.4 - Réactualisation du profil

Toute procédure de filtrage nécessite une phase d'apprentissage qui, si elle est bien menée, doit rapidement faire remonter l'efficacité du processus. Donc, à terme, le profil sera plus efficace qu'au départ, car les documents indésirables sont systématiquement ajoutés au profil ce qui permet de ne pas renouveler certaines erreurs. De même pour les documents les plus pertinents et dans le cas de niveaux de pertinence pour certains documents intermédiaires. Cette réactualisation du profil permet une auto-adaptation du processus de filtrage et donc un suivi de l'évolution du domaine ou des préoccupations de l'utilisateur. La procédure n'est donc pas figée dans le temps mais, au contraire, s'adapte aux fluctuations et aux ruptures dues souvent à la source elle même, au sujet, à l'utilisateur ou tout simplement à l'environnement (innovation, nouveaux acteurs, nouvelles applications, interfaçage avec d'autres domaines, évolution de la terminologie, maturité du sujet, …). Le principale difficulté est la participation active de l'utilisateur, qui doit réinjecter périodiquement dans le profil les documents les plus représentatifs soit de la variabilité du sujet, soit de nouvelles erreurs d'identification. De nouveaux termes peuvent aussi être élus et pris en compte dans la nouvelle définition du profil. Ce n'est pas une difficulté puisque le nombre de documents augment, le nombre de termes peut suivre. Bien sûr le système linéaire à résoudre est de plus en plus grand, mais ce n'est pas non plus un problème vu la puissance actuelle des processeurs. Le filtrage lui même ne nécessite que le calcul d'un produit scalaire par document, le plus difficile reste donc l'extraction des fréquences à la volée.

3.5 - Nouveau calcul de l'estimateur

Si la liste des termes retenus pour le modèle est reconduite, il suffit de déterminer la solution d'un nouveau système linéaire dont l'expression tient compte des nouveaux documents introduits dans le profil. Par contre, si la liste des termes est remise en cause (nouveaux mots-clés, nouvelles analyse du contenu sémantique), le nombre de paramètres va changer, l'ordre du système aussi. Dans les deux cas, il faut alors veiller à nettoyer le profil de documents devenus inutiles et il est recommandé d'effectuer une réévaluation de la pertinence du filtre (sur d'anciens document jugés) et éventuellement de réévaluer les documents qui avaient été préalablement rejetés et qui maintenant peuvent être acceptés par le nouveau profil. On peut ainsi récupérer des documents intéressants qui avaient été mal jugés précédemment (signaux faibles, émergences terminologiques, front de recherche) car le profil ne bénéficiait pas encore de la nuance sémantique du dernier apport.

4 - VALIDATION DE LA METHODE

4.1 - Base de tests utilisée

Nous avons utilisé une base de test très connue dans le monde du filtrage d'information : la base TREC. Nous avons généré plusieurs fichiers constitués d'environ 1800 documents déjà classés par les experts du domaine en documents pertinents et non pertinents. Pour représenter chaque document, nous ne gardons, dans un premier temps, que les 250 termes significatifs dont les fréquences sont les plus fortes Nous avons simulé, sur l'ensemble des documents de chaque série, l'évolution progressive d'un profil de filtrage. Au départ, le profil est initialisé

par les 300 premiers documents du fichier, dont une faible partie est constituée de documents pertinents. Un premier estimateur est calculé, il est ensuite appliqué soit à la totalité des documents (y compris ce qui ont servi à calculer l'estimateur), soit appliqué aux documents restants. L'efficacité du filtrage est alors évaluée par les indicateurs que nous allons décrire ci-dessous. Nous supposons que l'estimateur est recalculé dès qu'une série de 100 documents supplémentaire a été jugée. Pour les besoins du test, nous avons placé la même quantité de documents pertinents dans chaque tranche complémentaire de 100 documents. Nous pouvons donc connaître l'évolution des performances du filtrage sans que cette mesure dépende du nombre de nouveaux documents pertinents introduit à chaque fois dans le calcul de l'estimateur.

4.2 - Indicateurs utilisés

Dans les problèmes de filtrage, deux objectifs antagonistes sont à prendre en compte :

- Trouver un maximum de documents pertinents ou la part maximum
- Dans les documents récupérés la part de documents pertinents doit être maximum

Dans le premier cas, on cherche à obtenir le maximum d'information même s'il y a du bruit, dans le second cas, on ne veut pas de bruit dans les documents proposés quitte à en rater un nombre plus important que précédemment.

Pour atteindre simultanément ces deux objectifs, il faut donc réaliser un compromis et trouver un indicateur fiable permettant de le mesurer.

Nous allons utiliser les notations suivantes :

- Les documents pertinents (+)
- Les documents non pertinents (-)
- Les documents bien jugés (*V*)
- Les documents mal jugés (*F*)

Nous allons définir quatre mesures :

- Le nombre de vrais positifs, documents pertinents (+) bien jugés (*V*) : V_+
- Le nombre de vrais négatifs, documents non pertinents (-) bien jugés (*V*) : *V*-
- Le nombre de faux positifs, documents pertinents (+) mal jugés (*F*) : F_+
- Le nombre de faux négatifs, documents non pertinents (-) mal jugés (*F*) : *F*-

Le premier objectif est satisfait si :

Le taux de **rappel** $R_+ = V_+ /(V_+ + F_+)$ est maximum

Le second si :

Le taux de **précision** $P_+ = V_+ /(V_+ + V\text{-})$ est maximum

Une autre mesure utilisée est le taux de documents bien jugés soit :

$$\tau = (V_+ + V\text{-}) /(V_+ + V\text{-} + F_+ + F\text{-})$$

Dans notre cas qui est celui de la veille, l'objectif n'est pas de lire les documents filtrés mais de les analyser, il faut donc privilégier le taux de rappel. Par contre, si l'utilisateur final est lecteur et a donc une capacité de lecture limitée, il vaut mieux privilégier le taux de précision.
Pour valider notre démarche, nous avons défini un élément qui peut être paradoxal : le taux de rappel négatif. En l'associant au taux de rappel positif, nous obtenons une mesure du compromis cherché : une sorte de taux moyen.

$$R=\sqrt{R_+ R_-}=\sqrt{\frac{V_+ V_-}{(V_+ + F_+)(V_- + F_-)}}$$

En effet dans ce cas, les documents pertinents sont biens trouvés et les documents non pertinents bien écartés. Il faut bien entendu que ce taux s'approche de 1. Pour un filtrage aléatoire des documents (un sur deux) les valeur des différent indicateurs sont indiqués dans le tableau ci-dessous (le taux ***R*** est alors de 50%).

4.3 - Quelques résultats

Dans le cas de la base TREC, le nombre de documents pertinents est très inférieur au nombre de documents non pertinents. Nous avons donc constaté une disproportion flagrante dans la prise en compte par l'estimateur de ces deux population. Afin de rétablir l'équilibre, nous pondérons ces deux populations. La première idée est de simuler deux populations égales. Nous pouvons donc artificiellement gonfler la population des documents pertinents pour l'amener au niveau de celle des documents non pertinents, il suffit pour cela de multiplier chaque équation valant +1 par le coefficient suivant :

$$P = (F_+ + F-) / (V_+ + V-)$$

Les erreurs commises sur ces équations étant plus fortes, l'estimateur tient mieux compte des documents positifs, leur reconnaissance est ainsi améliorée, par contre le jugement sur les documents négatifs peut diminuer en qualité.

Comme cette pondération est un peu empirique, nous avons aussi essayé une sous pondération par la racine de P et une sur pondération par P à la puissance 3/2. Donc une série de quatre expériences :

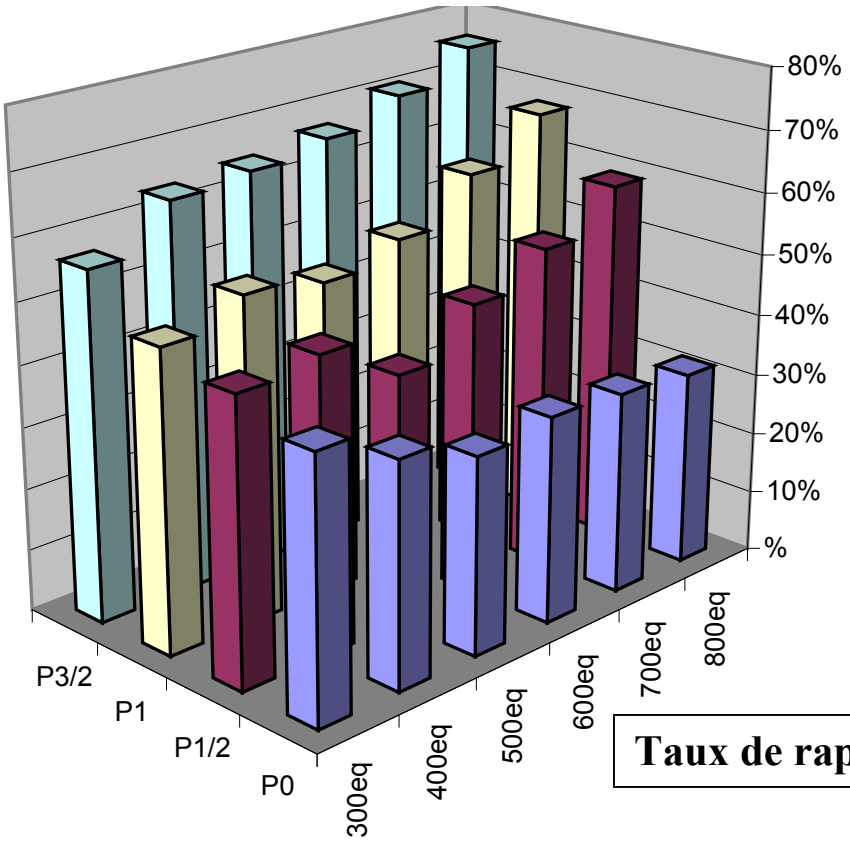
- P_0 sans pondération
- $P_{1/2}$ sous pondération
- P_1 pondération normale
- $P_{3/2}$ sur pondération

Les différents indices sont consignés dans les tableaux suivants :

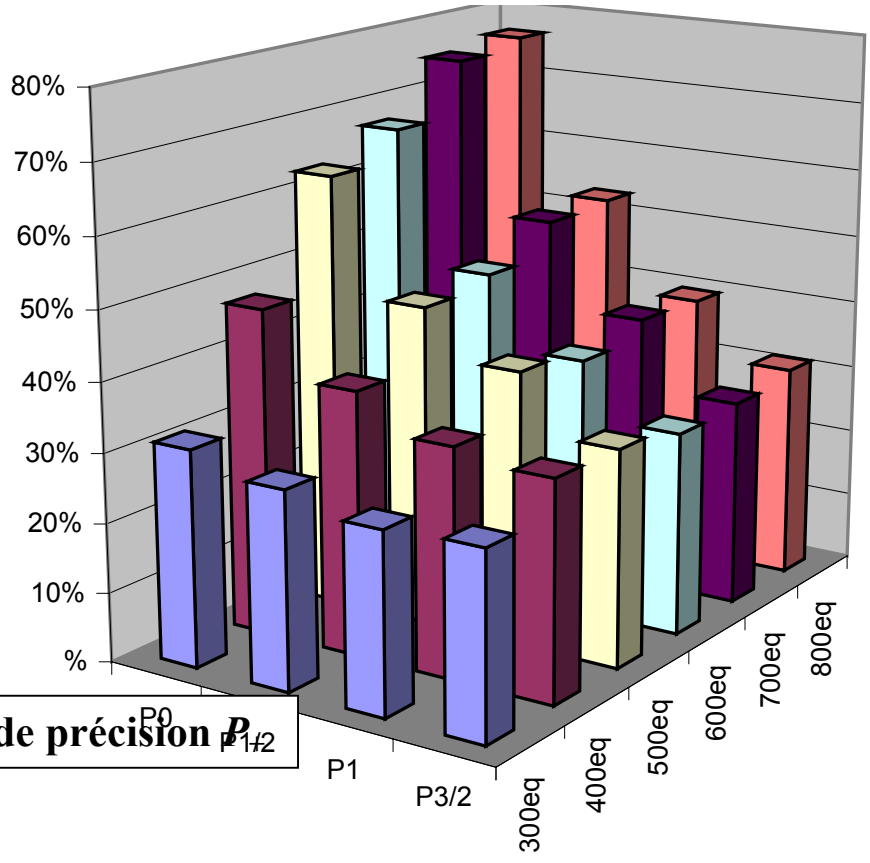
256		**P0**			**P 1/2**			**P 1**			**P 3/2**	
1805	21%	J+	J-	24%	J+	J-	27%	J+	J-	28%	J+	J-
300	M+	108	1311	M+	118	1253	M+	125	1194	M+	145	1151
	M-	238	148	M-	296	138	M-	355	131	M-	398	111
	31%	0,6	42%	29%	0,61	46%	26%	0,61	49%	27%	0,65	57%
	15%	J+	J-	18%	J+	J-	22%	J+	J-	25%	J+	J-
400	M+	93	1445	M+	122	1351	M+	135	1276	M+	163	1196
	M-	104	163	M-	198	134	M-	273	121	M-	353	93
	47%	0,58	36%	38%	0,64	48%	33%	0,66	53%	32%	0,7	64%
	12%	J+	J-	15%	J+	J-	18%	J+	J-	25%	J+	J-
500	M+	82	1501	M+	103	1430	M+	130	1351	M+	166	1186
	M-	48	174	M-	119	153	M-	198	126	M-	363	90
	63%	0,56	32%	46%	0,61	40%	40%	0,67	51%	31%	0,7	65%
	12%	J+	J-	15%	J+	J-	19%	J+	J-	27%	J+	J-
600	M+	86	1507	M+	121	1417	M+	138	1319	M+	171	1140
	M-	42	170	M-	132	135	M-	230	118	M-	409	85
	67%	0,57	34%	48%	0,66	47%	38%	0,68	54%	29%	0,7	67%
	11%	J+	J-	13%	J+	J-	19%	J+	J-	28%	J+	J-
700	M+	84	1521	M+	134	1428	M+	156	1315	M+	181	1126
	M-	28	172	M-	121	122	M-	234	100	M-	423	75
	75%	0,57	33%	53%	0,69	52%	40%	0,72	61%	30%	0,72	71%
	11%	J+	J-	13%	J+	J-	19%	J+	J-	27%	J+	J-
800	M+	81	1524	M+	151	1415	M+	173	1284	M+	194	1122
	M-	25	175	M-	134	105	M-	265	83	M-	427	62
	76%	0,56	32%	53%	0,73	59%	39%	0,75	68%	31%	0,74	76%
τ	50%	J+	J-	τ	**J+**	**J-**					**R**	
Aléa	M+	128	775	**M+**	**V+**	**V-**	**Avec intégration de l'estimateur**					
	M-	775	128	**M-**	**F-**	**F+**						
P+	14%	0,5	50%	**R+**								

Nous pouvons remarquer que les bons scores de précision sont obtenus pour un faible taux de rappel et inversement. Par contre le taux moyen maximum est obtenu pour des scores intermédiaires de ces deux indicateurs. L'erreur globale quant à elle suit le taux de précision car le nombre de documents pertinents est très faible.

Ci-dessous, les évolutions du taux de rappel et du taux de précision en fonction de la taille de l'échantillon (profil) et de la pondération corrigeant le déséquilibre entre documents pertinents et non pertinents pris en compte dans le calcul de l'estimateur.

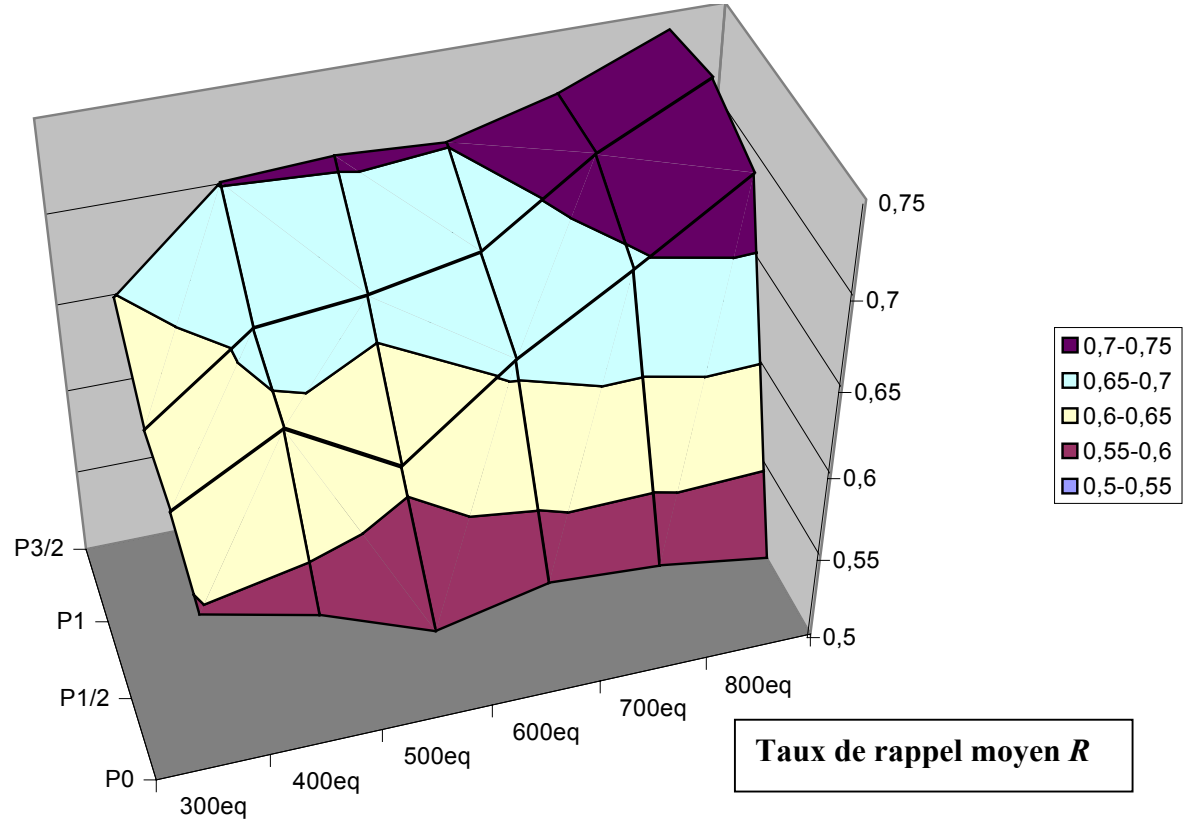


Taux de rappel R_+



Taux de précision P_+

Nous pouvons aussi constater que le taux moyen de réussite du filtrage augment avec le nombre de documents pris en compte du moment que ceux-ci sont pondérés pour égaliser exactement le poids des deux populations. Dans notre cas il n'y en en tout que 256 documents pertinents sur 1805, et dans le calcul des estimateurs leur nombre varie de 40 pour 300 documents à 90 pour 800 soit 10 de plus pour 100 nouveaux documents intégrant le profil.



4.4 - Intégration de l'outil

Pour l'instant, cet outil de filtrage fonctionne sur station SUN sous Solaris. Nous comptons prochainement l'introduire dans notre plate-forme Tétralogie (Unix et Linux) afin de nettoyer certains corpus du bruit dû à des équations de recherche trop larges (Internet, sujets difficiles à cerner, corpus généralistes, résultats de push en veille automatique). Une seconde utilisation peut être la recherche de sous corpus liés à un résultat d'analyse (production d'une équipe de recherche non isolée, ciblage d'un nouveau concept, d'une rupture, activité d'une communauté, environnement d'un sous domaine). En effet, ces entités sont un peu floues et ne peuvent pas être parfaitement identifiées par une simple liste de termes liés par une équation booléenne. L'idée est donc de partir de certains documents identifiés et d'en tirer toute la classe des documents voisins si possible triés par pertinence.

CONCLUSION

Parmi les différentes techniques de filtrage développées dans notre laboratoire (Vigie, Mercure), l'outils que nous proposons, basé sur l'algèbre linéaire, est plus conforme à la philosophie générale de notre plate-forme Tétralogie très orientée analyse de données à la française et donc très proche elle aussi de l'algèbre linéaire. Comme les premiers test sont particulièrement encourageants, nous comptons évaluer cet outil selon les mêmes conditions que les autres et offrir ainsi un alternative aux techniques statistiques et neuronales. Maintenant que le principe d'un estimateurs semble acquit, il nous faut améliorer et automatiser en partie les méthodes de sélection des termes les plus pertinents dans le cadre de cette approche. Faire évoluer le profil sera alors une tâche plus aisée, puisque le seul travail à la charge de l'utilisateur consistera à sélectionner de nouveaux documents pertinents en vue d'alimenter le profil. Les documents non pertinents quant-à eux pourront être sélectionnés pour entrer dans le profil en fonction des valeurs de leur fonction d'évaluation (les plus mauvais scores sont à privilégier afin de réaliser un recadrage efficace). Enfin, pour une utilisation sur la plate-forme, nous comptons

introduire une alimentation mixte du profil : dictionnaires de termes isolés (listes d'acteurs ou de mots-clés) et collection de documents repérés (externes ou issus du corpus à filtrer).

BIBLIOGRAPHIE

[DOUS99] B. Dousset, M. Salles

La Veille Scientifique par l'Analyse des Informations Ouvertes. 26th International Conference, Information Technologies in Science, Education and Business, (Yalta-Gurzuf, Crimée, Ukraine), may 17-30 1999.

[KARO99] S. Karouach, T. Dkaki, B. Dousset

Visualisation interactive de classifications d'informations. 8ièmes journées d'études sur les systèmes d'information élaborée: Bibliométrie - Informatique stratégique - Veille technologique, (Ile Rousse Corse France), CD-ROM, p. 45-61, 27 septembre-1° octobre 1999.

[ROUX99] C. Roux , B. Dousset

Une méthode de détection des signaux faibles: application à l'émergence des Dendrimères. Veille stratégique, scientifique et technologique : VSST'98, pp 349-357, (Toulouse, France), octobre 1998.

[DKAK00] T. Dkaki, B. Dousset, D. Egret, J. Mothe

Information discovery from semi-structured sources. Application to astronomical literature. Computer Physics Communications, Eds: Elsevier Science, V. 127 N° 2-3 , pp 198-206, 2000.

[DOUS00] B. Dousset, T. Dkaki, J. Mothe

Information mining in order to graphically summarize semi-structured document. 17th international CODATA Conference, (Baveno Italie), 15-19 octobre 2000.

[HUBE00] G. Hubert, J. Mothe, A. Benammar, T. Dkaki, B. Dousset, S. Karouach

Textual document Mining using graphical interface. International Human Computer Interaction, HCI International 2001 , New Orleans (USA). Lawence Erlbaum Associates - Publishers , Mahwah - New Jersey, pp 918-922 (volume 1), 05-10 août 2001.

[SALL00] M. Salles, Ph. Clermont, B. Dousset

MEDESIIE : une méthode de conception de systèmes d'intelligence écono-mique. IDMME'2000, (Montréal Canada), 16-19 mai 2000.

[BOUG01] M. Boughanem, B. Dousset

Relation entre le push adaptatif et l'optimisation des abonnements dans les centres de documentation. Veille stratégique, scientifique et technologique : VSST'01, pp 239-252, Vol 1, (Barcelone, Espagne), octobre 2001.

[KARO01] S. Karouach, B. Dousset

Visualisation interactive pour la découverte de connaissances. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 291-300, (Barcelone, Espagne), octobre 2001.

[MOTH01] J. Mothe, C. Chrisment, T. Dkaki, B. Dousset, D. Egret

Information mining: use of the document dimensions to analyse interactively a document set. 23rd BCS European Colloquium on IR Research: ECIR, Darmstadt. BCS IRSG, pp 66-77, 4-6 avril 2001.

[MULT01] J.-L. Multon, G. Lacombe, B. Dousset

Analyse bibliométrique des collaborations internationales de l'INRA. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 261-270, (Barcelone, Espagne), octobre 2001.

[SOSS01] D. Sosson, M. Vassard, B. Dousset

Portail pour la navigation en ligne dans les analyses stratégiques. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 347-358, (Barcelone, Espagne), octobre 2001.

[TMAR01] M. Tmar

Apprentissage incrémental dans un système de filtrage adaptatif. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 313-320, (Barcelone, Espagne), octobre 2001.

[DOUS02] B. Dousset, S. Karouach

Collaboration interactive entre classifications et cartes thématiques ou géographiques. 9ièmes rencontres de la société francophone de classification, (Toulouse France), 16-18 septembre 2002.