
Apprentissage et recherche documentaire : une approche probabiliste différentielle.

Benjamin Piwowarski

Laboratoire d'Informatique de Paris 6 (LIP6)
8, rue du Capitaine Scott
75015 Paris, France
Benjamin.Piwowarski@lip6.fr

RÉSUMÉ. Le but de la recherche documentaire (RD) consiste à trouver, parmi une base de documents, ceux qui répondent le mieux à une demande formulée par un utilisateur. Les expériences réalisées ont montré qu'il était impossible d'obtenir des résultats satisfaisants avec des systèmes où la représentation des documents ainsi que les paramètres utilisés lors de la recherche étaient figés. C'est pourquoi on emploie l'apprentissage pour modifier les paramètres du système de recherche, en se basant sur les jugements (feedback) que les utilisateurs peuvent porter sur la qualité des documents trouvés. Pourtant, l'application de telles techniques reste problématique. En effet, celles-ci sont soit complexes à mettre en œuvre (traitement off-line), soit difficiles à contrôler. Nous proposons une approche basée sur un modèle probabiliste de recherche documentaire qui permet d'utiliser le feedback de manière rapide et incrémentale. Dans cet article, nous étendons ce modèle pour ne plus regarder si un document répond à une requête dans l'absolu mais plutôt relativement à une base documentaire donnée. Ainsi, un jugement portant sur un seul document modifie l'ensemble du processus de recherche améliorant ainsi la rapidité de l'apprentissage.

ABSTRACT. The goal of information retrieval (IR) is to find, within a database of documents, those which satisfy a user's information need. Experiments carried out show that it is impossible to obtain perfect results with systems where both document representation and the parameters used for retrieving are fixed. Since user's relevance judgments are a source of evidence for information retrieval, learning from this feedback is an appealing idea. Many different learning techniques have successfully been used for relevance feedback. In most models, learning is either performed off line ; otherwise it is not well controlled. The approach we propose is based on a probabilistic model in which learning from feedback is simple and incremental. In this paper, we extend this model, allowing it to take into account feedback on the entire database while computing the score between a query and a single document. As a result, a judgment on a single document modifies the entire retrieval process, thus yielding on an improvement over learning.

MOTS-CLÉS: Recherche documentaire, apprentissage, modèles probabilistes.

KEYWORDS: Information Retrieval, learning, probabilistic model.

1. Introduction

La recherche documentaire (*RD*) a pour but de trouver, parmi un ensemble de documents, celui ou ceux qui répondent le mieux à une requête. Devant la complexité d'une telle tâche, la possibilité de concevoir des systèmes capables d'apprendre, grâce aux jugements formulés par les utilisateurs (*feedback*), à retrouver plus efficacement les documents pertinents est un enjeu de taille. Deux différents modes d'apprentissage existent en recherche documentaire. Le premier, *volatile*, permet d'améliorer progressivement le résultat d'une recherche, mais cet apprentissage ne dure que le temps d'une session avec l'utilisateur. Toute l'information apportée est perdue une fois la session terminée. Le second, *permanent*, utilise le feedback pour modifier le système à long terme. Dans cet article, nous nous focaliserons sur l'apprentissage permanent. L'organisation de l'article est la suivante. Dans la partie (2), les différentes approches utilisées en *RD* seront présentées et les différentes techniques d'apprentissage existantes seront détaillées. En (3), l'utilisation d'un nouveau modèle probabiliste sera proposée et justifiée. Les résultats obtenus seront présentés et commentés en (4). Le lecteur se référera aux annexes A, B et C pour les détails formels de l'approche.

2. Recherche documentaire et apprentissage

La plupart des approches utilisées en *RD* se basent sur une modélisation assez simple : un texte est représenté par un vecteur où chaque composante (binaire ou réelle) représente un terme (ou *concept*). Le principe sous-tendant toutes les différentes techniques utilisées en *RD* consiste à construire une fonction $S(q, d)$, appelée *mesure de similarité*, dont la valeur est d'autant plus grande que d répond à la requête q . Cette mesure permet alors d'ordonner les documents. Deux grands types d'approches issus des années 1970, qui ont aujourd'hui tendance à s'uniformiser [Spa 98], sont à distinguer : les approches probabilistes [CRE 98] et les approches vectorielles (*Vector Space Model*, *VSM*).

L'apprentissage en *RD* peut se résumer à la modification d'une des trois composantes d'un tel système : la représentation de la requête, la mesure de similarité et la représentation des documents. Keim et al [KEI 97] proposent un des rares modèles qui combine apprentissage volatile et permanent. Le modèle probabiliste sur lequel est basée leur approche utilise les jugements de pertinence de l'utilisateur afin d'affiner progressivement les résultats d'une session de recherche. Cette information est habituellement perdue pour toute nouvelle session, mais dans leur modèle, Keim et al. proposent d'altérer la représentation de toute nouvelle requête grâce aux requêtes passées – tout en modulant l'apport de toute ancienne requête en fonction de sa similarité avec la nouvelle. Bartell et al. [BAR 94, BAR 95, BAR 98] utilisent une méthode basée sur la descente de gradient pour optimiser la mesure de similarité dans un modèle vectoriel. Ils définissent une fonction différentiable, dont les paramètres sont ceux de la mesure de similarité, qui décroît lorsque l'ordre induit par la mesure de similarité se rapproche de l'ordre souhaité par les utilisateurs – le feedback définit ici un ordre de préférence des documents. Dans le cadre probabiliste, les méthodes de ré-

gression sont utilisées pour évaluer $P(R|q,d)$ à partir du feedback. Deux différentes techniques peuvent être employées, la régression *linéaire* [FUH 93, FUH 94] et la régression *logistique* [COO 93]. La technique de Brauen [BRA 71], *Document Vector Modification*, est une des plus simples utilisées pour l'apprentissage. Quand un document est jugé pertinent pour une requête, sa représentation est modifiée de manière à se «rapprocher» de la représentation de la requête. Bodoff [BOD 99] étend ce principe, mais évite une modification incontrôlée des documents : le but de l'apprentissage peut être vu comme la recherche, pour chaque document, d'un équilibre entre sa représentation initiale et la représentation des requêtes pour lesquelles il est pertinent.

Le modèle probabiliste sur lequel nous nous basons, appelé *Binary Independence Indexing Model* (BII), fait également partie des algorithmes qui modifient la représentation des documents grâce au feedback. Dans ce modèle, l'apprentissage est simple – il suffit de mettre à jour des statistiques, mais est par contre lent puisqu'un jugement n'influence que la représentation d'un document [CRE 98]. D'autres approches, comme celles de Belew [BEL 92a] et Kwok [KWO 92, KWO 94] proposent une approche connexionniste de la RD. Dans ces modèles, l'apprentissage est effectué grâce à des techniques spécifiques de mise à jours des poids du réseau.

3. L'approche probabiliste différentielle

Malgré la grande diversité des approches présentées, tous les systèmes montrent qu'il est possible pour un système de recherche documentaire de tirer partie de jugements de pertinence pour améliorer ses performances. Par contre, aucune de ces techniques n'offre à la fois un apprentissage incrémental (d'un point de vue pratique et non théorique), rapide et contrôlé (dans le sens où les modifications apportées au système ne sont pas seulement intuitives). Le modèle que nous présentons dans la suite, basé sur un modèle probabiliste, permet de combiner ces trois avantages.

Comme la plupart des approches probabilistes, notre modèle se fonde sur le *Probability Ranking Principle/PRP* [ROB 77] : $S(q,d)$ est construite de façon à ordonner de la même manière les documents que la probabilité $P(R|q,d)$ – probabilité qu'un document d soit pertinent pour une requête q . Dans notre modèle, tout document est considéré unique même si deux documents contiennent exactement les même termes. Un document est représenté par un ensemble de statistiques : pour tout terme c et tout document d , $R(c,d)$ (resp. $\overline{R}(c,d)$) représente le nombre de requêtes contenant c pour lesquelles d est pertinent (resp. n'est pas pertinent) ; $R(d)$ (resp. $\overline{R}(d)$) représente le nombre de requêtes pour lesquelles d est pertinent (resp. n'est pas pertinent). Par extension, nous noterons $R(c,\mathcal{D}')$ (resp. $R(\mathcal{D}')$) la somme des $R(c,d)$ (resp. $R(d)$) pour $d \in \mathcal{D}'$. Une requête est une conjonction d'événements simples, les Q_c (le terme c est présent dans la requête) et \overline{Q}_c (le terme c n'est pas présent dans la requête). Nous appellerons q cette représentation et utiliserons la notation $c \in q$ pour $Q_c \in q$.

D'un point de vue pratique, il faut établir quelques hypothèses permettant de simplifier suffisamment le problème pour qu'il puisse être traité. Une hypothèse courante en recherche documentaire, est de supposer une quasi indépendance des événements

simples (les Q_c et $\overline{Q_c}$) lorsque l'on connaît le document et la relation qui le lie à la requête (pertinence ou non-pertinence). Cette hypothèse est plus faible que l'hypothèse d'indépendance entre les termes, et est définie formellement de la façon suivante :

Hypothèse 1 (Linked dependence assumption) Pour tout $\mathcal{D}' \subseteq \mathcal{D}$ et toute requête q , nous avons la relation suivante :
$$\frac{P(\underline{q}|R, \mathcal{D}')}{P(\underline{q}|R, \mathcal{D}')} = \frac{\prod_{c \in \underline{q}} P(Q_c | \mathcal{D}', R) \prod_{c \notin \underline{q}} P(\overline{Q_c} | \mathcal{D}', R)}{\prod_{c \in \underline{q}} P(Q_c | \mathcal{D}', R) \prod_{c \notin \underline{q}} P(Q_c | \mathcal{D}', R)}$$

Une hypothèse supplémentaire permet d'ignorer, dans le calcul, les probabilités associées à tout terme qui n'est pas présent dans la requête.

Hypothèse 2 Pour toute requête q ,
$$\frac{P(\underline{q}|R, \mathcal{D}')}{P(\underline{q}|\overline{R}, \mathcal{D}')} = \frac{\prod_{c \in \underline{q}} P(Q_c | \mathcal{D}', R)}{\prod_{c \in \underline{q}} P(Q_c | \mathcal{D}', \overline{R})}$$

Ces deux hypothèses sont des *a priori* sur la recherche documentaire et la façon de modéliser l'ensemble des documents. On supposera de plus que la distribution des documents et que la probabilité qu'un document soit pertinent *a priori* sont uniformes :

Hypothèse 3 $\forall d, d' \in \mathcal{D}, d \neq d', P(d) = P(d')$ et $P(R|d) = P(R|d')$.

Contrairement à toutes les approches probabilistes utilisées en recherche documentaire, la méthode proposée ne cherche pas à estimer une probabilité pour un document d donné, mais cherche plutôt à l'estimer pour ce qu'on pourrait appeler un «anti-document» d (noté $\neg d$). D'une manière imagée, ce que nous cherchons à mesurer peut être résumé par la question suivante : «si on enlève le document d de la base documentaire, la probabilité d'y trouver un document pertinent est-elle moins élevée?». Si la réponse à cette question est non, alors ce document n'est pas pertinent. Dans le cas contraire, la variation de probabilité nous indiquera dans quelle mesure ce document est pertinent. La justification théorique de notre approche différentielle se base sur les calculs présentés en annexe. Nous proposons d'utiliser comme mesure de similarité la valeur $-\frac{P(R|\underline{q}, \neg d)}{P(\overline{R}|\underline{q}, \neg d)}$ et nous montrons en annexe B qu'elle respecte le *Probability Ranking Principle*. En annexe C, nous donnons une estimation de cette valeur. Nous montrons que l'ordre optimal de présentation des documents est l'ordre décroissant par rapport à la quantité :

$$S(q, d) = -|q| \log \frac{1 - \frac{\overline{R}(d)+2}{R(\mathcal{D})+2|\mathcal{D}|}}{1 - \frac{R(d)+2}{R(\mathcal{D})+2|\mathcal{D}|}} - \sum_{c \in \underline{q} \cap \underline{d}} \log \frac{1 - \frac{R(c, d)}{R(c, \mathcal{D})+|\mathcal{D}|-1}}{1 - \frac{\overline{R}(c, d)}{\overline{R}(c, \mathcal{D})+|\mathcal{D}|-1}}$$

En théorie, notre système pourrait s'amorcer sans aucune information (*i.e.* $R(c, d) = \overline{R}(c, d) = 0$ pour tout document d et pour tout terme c), et accumuler les jugements pour s'améliorer peu à peu. En pratique, une bonne initialisation de la représentation des documents est nécessaire. Pour cela, nous initialisons la représentation d'un document $d \in \mathcal{D}$ en «simulant» le traitement d'un nombre N de requêtes parmi lesquelles N_R ont pour bonne réponse d et $N_{\overline{R}}$ n'ont pas pour bonne réponse d . Cette initialisation reste très simple puisque les seules données nécessaires sont contenues dans le document. La méthode dont la validation reste expérimentale, consiste à initialiser pour chaque terme c d'un document d les valeurs

$R(c,d)$ et $\bar{R}(c,d)$ de la façon suivante: $R(c,d) = \left\lfloor 1 + (N_R - 1) \log \frac{N_c(d)+1}{\max_d + 1} \right\rfloor$ et $\bar{R}(c,d) = \left\lfloor 1 + (N_{\bar{R}} - 1) \left(1 - \frac{\log(N_c(d)+1)}{\log(\max_d + 1)} \right) \right\rfloor$, où $\lfloor x \rfloor$ représente la partie entière de x , $N_c(d)$ le nombre d'occurrences du terme c dans d , et \max_d le maximum du nombre d'occurrences du terme dans d .

4. Résultats obtenus

Pour l'évaluation, deux différentes collections ont été utilisées (http://www.dcs.gla.ac.uk/ir_resources). La première est la collection Cranfield (1398 documents et 1837 jugements de pertinence). La seconde est la collection CISI (1460 documents et 3114 jugements de pertinence). Les algorithmes évalués sont :

- *TF-IDF* : algorithme utilisé par SMART [SAL 71]. Utilisation de TF-IDF pour le calcul de φ et du *cosinus* comme mesure de similarité.

- *DIFF_n* avec $n = 0,1,10$: notre méthode différentielle. Trois différentes conditions expérimentales ont été évaluées. Pour la première ($n = 0$), il n'y a pas d'apprentissage. Dans la seconde ($n = 1$), les requêtes du groupe d'apprentissage sont utilisées pour modifier la représentation des documents. Quant à la troisième ($n = 10$), chaque requête du groupe d'apprentissage simule l'apprentissage à partir de dix requêtes avec le même texte et les mêmes documents pertinents.

Pour chaque collection, les requêtes ont été divisées en deux groupes différents (tableau 1), le premier utilisé pour l'apprentissage (pour *DIFF_n*, $n > 0$), le second pour l'évaluation. Les performances des algorithmes sont résumées par plusieurs mesures classiques en recherche documentaire. La courbe rappel/précision (figures 1 et 2) montre pour chaque taux de rappel x (rapport entre le nombre de réponses trouvées et le nombre de réponses pertinentes), la précision (rapport entre le nombre de réponses pertinentes et le nombre de réponses trouvées) maximum atteinte pour un taux de rappel d'au moins x . La précision à n documents (tableau 2) donne le nombre moyen de documents pertinents trouvés par l'algorithme parmi les n premiers.

	Cranfield		CISI	
	App.	Eval.	App.	Eval.
Nombre de requêtes	158	67	67	44
Nombre moyen de jugements par requête	8.4	7.5	27.3	29.3
Nombre moyen de jugements par document	0.95	0.36	1.25	0.88

TAB. 1 – Caractéristiques des jeux de requêtes

Les résultats obtenus sont encourageants lorsque l'on considère ces observations. Quand aucun jugement de pertinence n'est utilisé (*DIFF₀*), les performances de notre modèle sont proches de celles de SMART (figures 1 et 2 à gauche et tableau 2). Ce

Nombre moyen de documents pertinents pour..	Cranfield				CISI		
	TF-IDF	DIFF ₀	DIFF ₁	DIFF ₁₀	TF-IDF	DIFF ₀	DIFF ₁
1 document	0.66	0.69	0.72	0.73	0.32	0.42	0.42
2 documents	1.12	1.13	1.15	1.21	0.58	0.71	0.77
3 documents	1.52	1.49	1.54	1.63	1.00	1.00	1.03
5 documents	2.01	1.97	2.03	2.23	1.71	1.65	1.78
10 documents	2.73	2.72	2.82	3.16	2.90	2.84	2.94

TAB. 2 –. Précision à n documents. Les meilleures performances sont en gras.

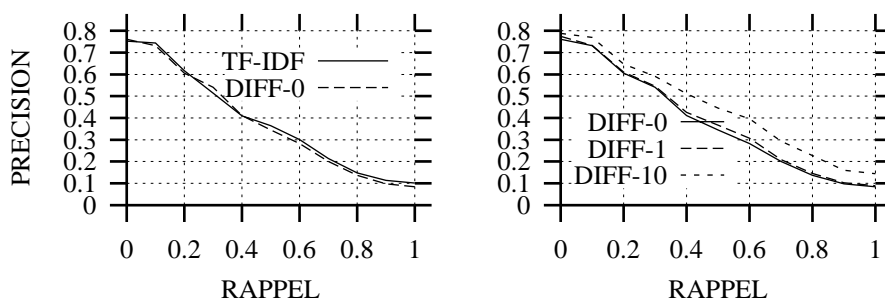


FIG. 1 –. Cranfield 1400 : Comparaison avec SMART / Effet de l'apprentissage

résultat est d'autant plus important que notre mesure de similarité est totalement nouvelle et n'est comparable directement à aucune autre, alors que SMART est un algorithme éprouvé. De plus, l'apprentissage améliore les résultats obtenus (figures 1 et 2 à droite et tableau 2). L'effet est d'autant plus grand que le nombre de requêtes simulées est grand ($DIFF_{10}$). Cette observation suggère l'idée que notre modèle pourrait tirer partie de l'utilisation de différents niveaux de pertinence (très pertinent, moyennement

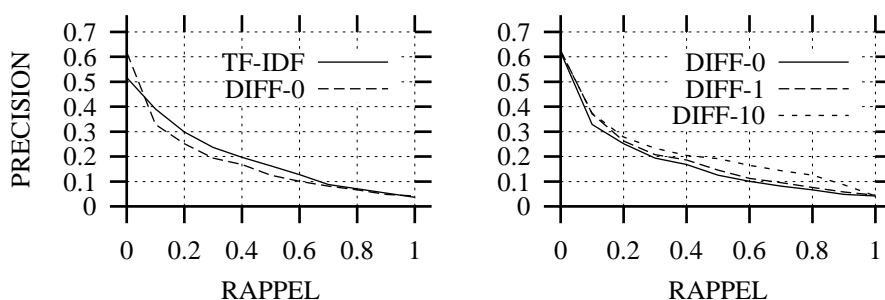


FIG. 2 –. CISI : Comparaison avec SMART / Effet de l'apprentissage

pertinent, etc.) en assignant à chaque niveau une constante représentant le nombre de fois où le triplet (requête, document, pertinence) doit être simulé.

5. Conclusion et perspectives

Dans cet article, un nouveau modèle probabiliste de recherche documentaire a été présenté. Celui-ci permet de construire un système de RD capable d'apprendre de manière simple et rapide grâce aux jugements de pertinence des utilisateurs. De plus, ses performances sans apprentissage sont comparables à celles d'un algorithme classique. Comme ce modèle remplace progressivement la représentation initiale d'un document par celle des requêtes pour lesquelles il est (ou n'est pas) pertinent, une analyse linguistique poussée de la requête permettant de dégager des informations de haut niveau – comme par exemple la nature de la question : où, comment, etc. – pourrait compléter de manière efficace un tel modèle. Parmi les autres améliorations possibles, l'utilisation d'une base structurée permettrait d'accélérer les recherches et l'apprentissage en regroupant les caractéristiques communes à des sous-ensembles de documents.

Remerciements

Je voudrais remercier la société LexiQuest (<http://www.lexiquest.com>) où tout ce travail a été effectué. Un grand merci à Roger Le Borgne (LexiQuest) et Patrick Gallinari (LIP6) qui ont supervisé ce travail.

A. Définitions et notations

Soit $\Omega = \mathcal{Q} \times \mathcal{D}$ l'univers de probabilité où \mathcal{Q} est l'ensemble des requêtes et \mathcal{D} l'ensemble des documents. Les événements utilisés dans la suite sont les suivants :

- «avoir le document d », $d = \{(q_i, d_j) \in \mathcal{Q} \times \mathcal{D} | d = d_j\}$ et l'événement avoir un document de $\mathcal{D}' \subset \mathcal{D}$, noté $\mathcal{D}' = \bigcup_{d \in \mathcal{D}'} E(d)$.
- «avoir la requête q », noté q ou $E(q) = \{(q_i, d_j) \in \mathcal{Q} \times \mathcal{D} | q = q_i\}$.
- «pertinence», $R = \{(q_i, d_j) \in \mathcal{Q} \times \mathcal{D} | d_j \text{ est pertinent pour } q_i\}$.
- L'événement la concept c est présent dans la requête, noté Q_c ou $E(Q_c) = \{(q_i, d_j) \in \mathcal{Q} \times \mathcal{D} | c \in q_i\}$.

B. Relation document-antidocument

La définition de l'événement d implique, avec $\neg d = E(\mathcal{D} \setminus \{d\})$:

$$P(R|\underline{q}, \neg d) = \frac{P(R|\underline{q})}{P(\neg d|\underline{q})} \left(1 - \frac{P(d|\underline{q})}{P(R|\underline{q})} P(R|d, \underline{q}) \right) \text{ avec Bayes}$$

$$\text{et donc } \frac{P(R|\underline{q}, \neg d)}{P(\overline{R}|\underline{q}, \neg d)} = \frac{P(R|\underline{q}) - P(d)P(R|d, \underline{q})}{(1 - P(R|\underline{q})) - P(d)(1 - P(R|d, \underline{q}))} \quad [1]$$

La formule (1) nous permet de dire $-\frac{P(R|\underline{q}, \neg d)}{P(\overline{R}|\underline{q}, \neg d)}$ est une fonction strictement croissante de $P(R|\underline{q}, d)$. Ceci nous permet de créer un ordre qui respecte le *Probability Ranking Principle* en les ordonnant les documents en fonction de $-\frac{P(R|\underline{q}, \neg d)}{P(\overline{R}|\underline{q}, \neg d)}$. Puis,

$$\begin{aligned} \log \frac{P(R|\underline{q}, \neg d)}{P(\overline{R}|\underline{q}, \neg d)} &= \log \frac{P(R|\neg d)}{P(\overline{R}|\neg d)} + \log \frac{P(\underline{q}|R, \neg d)}{P(\underline{q}|\overline{R}, \neg d)} \text{ avec Bayes} \\ &= \log \frac{P(R)}{P(\overline{R})} + \sum_{c \in \underline{q}} \log \frac{P(Q_c|R, \neg d)}{P(Q_c|\overline{R}, \neg d)} \text{ (hypothèses 1, 2 et 3)[2]} \end{aligned}$$

C. Estimation de probabilité

La probabilité qu'un terme c soit présent dans une requête pour qui le document d est pertinent est $p_{c,d} = P(Q_c \wedge d \wedge R)$. Soit $X_{c,d}^i$ une variable aléatoire qui prend la valeur 1 quand $E(Q_c) \wedge E(d) \wedge E(R)$ est vrai, et 0 autrement. $X_{c,d}^i$ suit alors une loi de BERNOUILLI de paramètre $p_{c,d}$. Soit $Y_{c,d} = X_{c,d}^1 + \dots + X_{c,d}^{N_q}$ une variable aléatoire qui représente le nombre de fois où un document d est pertinent pour une requête qui contient le concept c (N_q représente le nombre de requêtes pour lesquelles nous avons un jugement de pertinence). Nous pouvons facilement faire l'hypothèse que $X_{c,d}^i$ et $X_{c,d}^j$, $i \neq j$ sont indépendantes. Par conséquent, $Y_{c,d}$ suit une loi binomiale de paramètres $p_{c,d}$ et N_q . De plus, $P(Y_{c,d} = k) = C_{N_q}^k p_{c,d}^k (1 - p_{c,d})^{N_q - k}$ avec $k = 0, \dots, N_q$ et $\mathbb{E}(Y_{c,d}) = N_q p_{c,d}$. En posant $Y_{c, \mathcal{D}'} = \sum_{d \in \mathcal{D}'} Y_{c,d}$ nous obtenons:

$$\begin{aligned} \mathbb{E}(Y_{c, \mathcal{D}'}) &= N_q P(R) P(\mathcal{D}'|R) P(Q_c|\mathcal{D}', R) \approx R^*(c, \neg d) \\ \text{et } \frac{\mathbb{E}(Y_{c, \neg d})}{\mathbb{E}(Y_{c, \mathcal{D}'})} &= P(\neg d) \frac{P(Q_c|\neg d, R)}{P(Q_c|R)} \approx \frac{R^*(c, \neg d)}{R^*(c, \mathcal{D}')} \quad [3] \end{aligned}$$

où $R^*(c, \neg d)$ et $\overline{R}^*(c, d)$ représentent la valeur estimée de $R(c, \neg d)$ et $\overline{R}(c, d)$ si pour chaque requête passée et chaque document nous connaissions les jugements de pertinence. Pour obtenir une estimation de $R^*(c, \mathcal{D}')$, nous devons prendre en compte l'hypothèse 3: un document doit être pertinent pour un même nombre de requête qu'un autre. Notons R_q celui-ci. De plus, on suppose que pour tout document d et tout terme c quatre jugements sont connus représentant les différents cas de figure possible: le document est (n'est pas) pertinent pour la requête et la requête contient (ne contient pas) le terme c . Cette initialisation représente donc une incertitude totale, et

est directement incluse dans la formule de similarité, pour des raisons évidentes d'optimisation. La valeur $\frac{R(c, \mathcal{D}') + |\mathcal{D}'|}{R(\mathcal{D}') + 2|\mathcal{D}'|}$ est le pourcentage du nombre de couples requête-document liés par la relation de pertinence et dont le document est dans \mathcal{D}' pour qui la requête contient le concept c . Ce qui nous amène à établir l'approximation suivante de $R^*(c, \mathcal{D}')$:

$$R^*(c, \mathcal{D}') \approx \frac{R(c, \mathcal{D}') + |\mathcal{D}'|}{R(\mathcal{D}') + 2|\mathcal{D}'|} \times R_q \times |\mathcal{D}'|$$

et, avec la formule (3) et les relations $R(c, \neg d) = R(c, \mathcal{D}) - R(c, d)$ et $R(\neg d) = R(\mathcal{D}) - R(d)$, nous obtenons une estimation de $\tilde{P}(Q_c | \neg d, R)$:

$$\tilde{P}(Q_c | \neg d, R) = \frac{R^*(c, \neg d)}{R^*(c, \mathcal{D})} \times \frac{P(Q_c | R)}{P(\neg d)} = \frac{1 - \frac{R(c, d) + 1}{R(c, \mathcal{D}) + |\mathcal{D}|}}{1 - \frac{R(d) + 2}{R(\mathcal{D}) + 2|\mathcal{D}|}} \times \frac{|\mathcal{D}| - 1}{|\mathcal{D}|} \times \frac{P(Q_c | R)}{P(\neg d)}$$

Un résultat analogue peut être trouvé pour $P(Q_c | \neg d, \bar{R})$ et donc :

$$\frac{\tilde{P}(Q_c | \neg d, R)}{\tilde{P}(Q_c | \neg d, \bar{R})} = \frac{1 - \frac{R(c, d) + 1}{R(c, \mathcal{D}) + |\mathcal{D}|}}{1 - \frac{\bar{R}(c, d) + 1}{\bar{R}(c, \mathcal{D}) + |\mathcal{D}|}} \times \frac{1 - \frac{\bar{R}(d) + 2}{\bar{R}(\mathcal{D}) + 2|\mathcal{D}|}}{1 - \frac{R(d) + 2}{R(\mathcal{D}) + 2|\mathcal{D}|}} \times \frac{P(Q_c | R)}{P(Q_c | \bar{R})} \quad [4]$$

Grâce à ce dernier résultat, nous pouvons écrire une approximation de la formule (1), en notant $c \in \underline{d} \iff R(c, d) \neq 0$ ou $\bar{R}(c, d) \neq 0$:

$$\begin{aligned} \log \frac{\tilde{P}(R | \underline{q}, \neg d)}{\tilde{P}(\bar{R} | \underline{q}, \neg d)} &= \log \frac{P(R)}{P(\bar{R})} + \sum_{c \in \underline{q}} \log \frac{\left(1 - \frac{1}{R(c, \mathcal{D}) + |\mathcal{D}|}\right) P(Q_c | R)}{\left(1 - \frac{1}{\bar{R}(c, \mathcal{D}) + |\mathcal{D}|}\right) P(Q_c | \bar{R})} \\ &+ |\underline{q}| \log \frac{1 - \frac{\bar{R}(d) + 2}{\bar{R}(\mathcal{D}) + 2|\mathcal{D}|}}{1 - \frac{R(d) + 2}{R(\mathcal{D}) + 2|\mathcal{D}|}} + \sum_{c \in \underline{q} \cap \underline{d}} \log \frac{1 - \frac{R(c, d)}{R(c, \mathcal{D}) + |\mathcal{D}| - 1}}{1 - \frac{\bar{R}(c, d)}{\bar{R}(c, \mathcal{D}) + |\mathcal{D}| - 1}} \quad [5] \end{aligned}$$

où les deux premiers termes sont constants pour une requête donnée et ne modifient donc pas l'ordre obtenu : ils peuvent donc être ignorés.

D. Bibliographie

- [BAR 94] BARTELL B. T., COTTRELL G. W., BELEW R. K., « Optimizing Parameters in a Ranked Retrieval System Using Multi-Query Relevance Feedback », *Proceedings Symposium on Document Analysis and Information Retrieval*, Las Vegas, Nevada, avril 1994, University of Nevada.
- [BAR 95] BARTELL B., COTTRELL G. W., BELEW R., « Learning to Retrieve Information », NIKLASSON L., BODEN M., Eds., *Current trends in connectionism: Proceedings of the Swedish Conference on Connectionism*, LEA: Hillsdale, 1995.
- [BAR 98] BARTELL B., COTTRELL G. W., BELEW R., « Optimizing Similarity using Multi-Query Relevance Feedback », *Journal of the American Society for Information Science*, vol. 49, n° 8, 1998, p. 742-761.

- [BEL 92a] BELEW R. K., « Adaptive Information Retrieval: Using a connectionist representation to retrieve and learn about documents », Belkin, van Rijsbergen [BEL 92b], p. 11-20.
- [BEL 92b] BELKIN N. J., VAN RIJSBERGEN C. J., Eds., *Proceedings of the ACM SIGIR 12th International Conference on Research and Development in Information Retrieval*, Cambridge, Massachusetts, USA, juin 1992, ACM Press.
- [BOD 99] BODOFF D., « A Re-Unification of Two Competing Models for Document Retrieval », *Journal of the American Society for Information Science*, vol. 50, n° 1, 1999, p. 49-64.
- [BRA 71] BRAUEN T., « Document Vector Modification », Salton [SAL 71], Chapitre 24, p. 456-484.
- [COO 93] COOPER W. S., CHEN A., GEY F. C., « Full Text Retrieval based on Probabilistic Equations with Coefficients fitted by Logistic Regression », Harman [HAR 93], p. 57-66.
- [CRE 98] CRESTANI F., LALMAS M., VAN RIJSBERGEN C. J., CAMPBELL I., « "Is this document Relevant?... Probably": a survey of Probabilistic Models in Information Retrieval », *ACM Computing surveys*, vol. 30, n° 4, 1998, p. 528-552.
- [FUH 93] FUHR N., PFEIFER U., BREMKAMP C., POLLMANN M., BUCKLEY C., « Probabilistic Learning Approaches for Indexing and Retrieval with the TREC-2 Collection », Harman [HAR 93], p. 67-74.
- [FUH 94] FUHR N., PFEIFER U., « Probabilistic Information Retrieval as a Combination of Abstraction, Inductive Learning, and Probabilistic Assumptions », *ACM Transactions On Information Systems*, vol. 12, n° 1, 1994, p. 92-115.
- [HAR 93] HARMAN D. K., Ed., *NIST Special Publication 500-215: The Second Text REtrieval Conference (TREC-2)*, Gaithersburg, MD, août-septembre 1993, Department of Commerce, National Institute of Standards and Technology.
- [KEI 97] KEIM M., LEWIS D. D., MADIGAN D., « Bayesian Information Retrieval: Preliminary Evaluation », MADIGAN D., SMYTH P., Eds., *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, Ft. Lauderdale, Florida, janvier 1997, p. 303-310.
- [KWO 92] KWOK K., « A Neural Network for Probabilistic Information Retrieval », Belkin, van Rijsbergen [BEL 92b], p. 11-20.
- [KWO 94] KWOK K. L., GRUNFELD L., LEWIS D. D., « TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS », HARMAN D. K., Ed., *NIST Special Publication 500-226: Overview of the Third Text REtrieval Conference (TREC-3)*, Gaithersburg, MD, novembre 1994, U. S. Dept. of Commerce, National Institute of Standards and Technology, p. 247-255.
- [ROB 77] ROBERTSON S. E., « The probability ranking principle in IR », *Journal of Documentation*, vol. 33, 1977, p. 294-304.
- [SAL 71] SALTON G., Ed., *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971.
- [Spa 98] SPARK JONES K., WALKER S., ROBERTSON S., « A probabilistic model of information retrieval: development and status », rapport n° 446, août 1998, Computer Laboratory, University of Cambridge.

Annexe pour le service de fabrication

Article pour les actes :

CAP'2000

Auteurs :

Benjamin Piwowarski

Titre de l'article :

*Apprentissage et recherche documentaire :
une approche probabiliste différentielle.*

Titre abrégé :

Apprentissage et recherche documentaire

Traduction du titre :

Learning in Information Retrieval: a Probabilistic Differential Approach

Date de cette version :

4th April 2001

Coordonnées des auteurs :

- téléphone : (33) 1 44 27 74 91
- télécopie : (33) 1 44 27 70 00
- Email : Benjamin.Piwowarski@lip6.fr

Logiciel utilisé pour la préparation de cet article :

\LaTeX , avec le fichier de style `article-hermes.cls`,
version 1.4 du 26/01/2000.

Formulaire de copyright :

Joindre le formulaire de copyright signé, récupéré sur le web à l'adresse
<http://www.hermes-science.com>