Extraction de la localisation des termes pour le classement des documents

Annabelle MERCIER*, Michel BEIGBEDER*

*École des Mines de Saint-Etienne
158 cours Fauriel F 42023 Saint-Étienne Cedex 2 FRANCE
mercier,beigbeder@emse.fr

Résumé. Trouver et classer les documents pertinents par rapport à une requête est fondamental dans le domaine de la recherche d'information. Notre étude repose sur la localisation des termes dans les documents. Nous posons l'hypothèse que plus les occurrences des termes d'une requête se retrouvent proches dans un document alors plus ce dernier doit être positionné en tête de la liste de réponses. Nous présentons deux variantes de notre modèle à zone d'influence, la première est basée sur une notion de proximité floue et la seconde sur une notion de pertinence locale.

1 Introduction

Le domaine de la recherche d'information, bien connu à travers les moteurs de recherche sur le Web, utilise différents modèles. Ces derniers précisent comment sélectionner et ordonner les documents qui répondent aux besoins d'informations des utilisateurs. Il en existe principalement trois familles (Baeza-Yates et Ribeiro-Neto, 1999) : (a) les modèles ensemblistes (booléen, à ensembles flous et booléens étendus), (b) les modèles algébriques (vectoriel et indexation sémantique latente) et (c) les modèles probabilistes (basés sur les réseaux d'inférence, les réseaux bayésiens et les réseaux de croyance). Notre modèle est basé non seulement sur les familles de modèle ensemblistes et algébriques, mais aussi sur une des premières idées fondatrice de la recherche d'information formulée par Luhn (Luhn, 1958) qui consiste à s'appuyer d'une part, sur la fréquence des termes et d'autre part sur la position relative des termes de la requête dans les documents. Le premier aspect relatif à l'utilisation de la fréquence des termes a été beaucoup développé dans le cadre des modèles algébriques, par contre, le second concernant la proximité entre les occurrences des termes n'a reçu que peu d'attention, notre étude permet d'approfondir ce dernier point.

Tout d'abord, nous rappelons certains modèles classiques ainsi que les quelques méthodes qui utilisent la proximité. Ensuite, nous présentons les deux variantes de notre modèle à zone d'influence avant de conclure.

2 État de l'art

La méthode d'indexation associée à un modèle de recherche d'information permet de construire les représentants des documents et s'appuie généralement sur les occurrences des termes trouvés dans les documents. Nous notons T l'ensemble des termes et D celui des documents.

Dans le modèle booléen classique, un document est représenté par l'ensemble des termes qui le composent et, une requête est formulée à l'aide d'une expression booléenne représentée par un arbre où les feuilles sont des termes et les nœuds sont les opérateurs ET et OU. Le score attribué aux documents est pris dans l'ensemble $\{0, 1\}$ et ces derniers ne peuvent donc pas être classés ce qui est un inconvénient majeur. Néanmoins, l'une des forces de ce modèle est l'expressivité du langage de requête.

Le modèle vectoriel correspond bien à la première idée de Luhn car la fréquence des termes est prise en compte pour attribuer un score aux documents. Le poids $w(d, t)$ du terme t dans le document d dépend de façon croissante de la fréquence du terme dans ce document et de façon décroissante de la fréquence documentaire de ce terme. Un document (resp. une requête) est représenté par un vecteur et la valeur de similarité entre un document et une requête est le plus souvent calculée avec la méthode du cosinus. Le modèle de requête qui est un sac de mots, est donc plus simple mais moins expressif que celui du modèle booléen. Le modèle vectoriel possède l'avantage de ranger les documents qui peuvent être présentés à l'utilisateur par ordre décroissant de pertinence calculée par le système. La possibilité de classer les documents, inexistante dans le modèle booléen, est fondamentale car elle est à la base des méthodes d'évaluation des systèmes. Son absence dans le modèle booléen classique a conduit à l'introduction des modèles booléens étendus et à l'utilisation des ensembles flous.

Pour graduer le score dans le cadre des modèles ensemblistes, plusieurs modèles basés sur la théorie des sous ensembles flous ont été développés (Miyamoto, 1990). A chaque terme $t \in T$ est associée une fonction μ_t qui traduit le degré d'appartenance d'un document à l'ensemble flou correspondant au terme. Une requête est aussi représentée par un arbre et un nœud avec l'opérateur OU (resp. ET) est évalué en prenant le maximum (resp. minimum) sur les valeurs de ses fils, ce qui correspond à la réunion (resp. intersection) floue des sous-ensembles flous correspondant à ses fils. Finalement, pour une requête donnée, le score d'un document est pris dans l'intervalle [0,1], ce qui permet contrairement au modèle booléen classique de classer les documents.

Une extension du modèle booléen consiste à ajouter un opérateur de proximité au langage de requête pour exprimer la position relative entre deux termes. Il est souvent nommé NEAR, ADJ-acent, ou WINDOW (Salton et al., 1983) et se comporte comme un ET avec une contrainte supplémentaire qui permet de préciser une distance maximale entre deux occurrences de termes, comme dans A NEAR 5 B. Dans notre modèle, nous n'ajoutons pas l'opérateur NEAR, car ce dernier ne peut s'appliquer qu'à des feuilles et sa généralisation aux sous-arbres est inconsistante (Mitchell, 1973). Cependant, la pertinence reste binaire donc aucun classement n'est envisageable.

D'autres approches utilisent directement la proximité des termes (Clarke et al., 2000; Hawking et Thistlewaite, 1995; Rasolofo et Savoy, 2003) en recherchant dans le texte les intervalles contenant les mots de la requête. Une contribution au score est calculée pour chaque intervalle (plus il est court, plus le score est élevé), et finalement le score d'un document dépend de la somme de ces contributions. Le phase de sélection des intervalles ainsi que le calcul de la contribution de chaque intervalle sont différents selon les méthodes. Nous avons déjà effectué une étude comparative qui présente en détails chacune de ces méthodes (Mercier, 2004). Les résultats obtenus pour ces méthodes sont meilleurs que ceux obtenus pour les modèles traditionnels.

3 Notre modèle

En prenant en compte soit l'appartenance soit la fréquence d'un terme dans un document, les modèles booléen et vectoriel procèdent avec une approche **globale** de l'influence des occurrences d'un terme sur la pertinence d'un document à une requête. C'est-à-dire que quelque soient les positions des occurrences d'un terme, cela n'a pas de conséquence sur le score de pertinence. Cependant, le sens du texte dans un document ne dépend pas seulement du vocabulaire employé mais aussi de l'agencement des termes de ce vocabulaire. Notre approche est **locale** dans le sens où nous modélisons une *influence* des occurrences. Cette influence est soit :

- **une proximité au terme** : en un endroit du texte, est-on proche d'une occurrence de ce terme ? Cette proximité sera graduée, et nous emploierons le terme de *proximité floue* ;
- **une pertinence locale** : un endroit du texte est-il pertinent à un terme t ? Cette pertinence sera d'autant plus élevée qu'il y a de nombreuses occurrences de ce terme à proximité.

Dans les deux cas, l'influence d'une occurence d'un mot est représentée à l'aide d'une *fonction d'influence*. Nous appelons ainsi une fonction définie sur $\mathbb{R}$, à support borné, prenant ses valeurs dans $[0, 1]$, croissante sur $\mathbb{R}^-$, et décroissante sur $\mathbb{R}^+$. Différentes fonctions d'influence peuvent être utilisées : il est d'abord possible de choisir une famille de fonctions (fonctions de Hamming, de Hanning, gaussiennes, rectangulaires, triangulaires, etc.), ensuite, des valeurs différentes peuvent être fixées pour les paramètres de ces fonctions afin d'obtenir une fonction d'influence différente pour chaque terme de la requête. En particulier, nous appelons k le paramètre qui contrôle la largeur de la zone d'influence. Pour une occurrence d'un terme, la translation $g(x) = f(x - i)$ d'une fonction d'influence f sert à modéliser la proximité floue (resp. la pertinence locale au terme qui a une occurence à la position x). Par exemple, pour une fonction triangulaire, la valeur au point x est égale à 1 puis décroit de $\frac{1}{k}$ aux positions voisines jusqu'à atteindre la valeur 0. Nous pouvons exprimer cette fonction d'influence par $f(x) = \max(\frac{k-|x|}{k}, 0)$.

3.1 Modèle à proximité floue

Il est naturel de considérer que la valeur de la proximité floue à un terme t en une position x d'un document est celle de la plus proche occurrence du terme t. Par exemple, pour un terme qui apparaît aux positions $x = 2$ et $x = 5$, la valeur de proximite floue à la position $x = 3$ est la valeur maximale entre celles de ces deux occurences, soit la proximité de la plus proche, c'est-à-dire celle de l'occurrence du terme en $x = 2$. Comme les fonctions d'influence précédement définies sont décroissantes par rapport à la distance des occurrences, en une position x du texte cela revient à prendre la valeur de proximité floue maximale et on peut poser $p_t^d(x) = \max_{i \in Occ(t,d)} f(x - i)$ où $Occ(t, d)$ est l'ensemble des positions des occurrences du terme t dans le document d et f la fonction d'influence choisie. Les feuilles de l'arbre de la requête portent donc des fonctions de proximité correspondant aux termes. Par exemple, la fonction p_A (resp. p_B) associe la valeur de proximité floue au terme A (resp. B) à toutes les positions d'un document d. Nous généralisons maintenant ces fonctions sur les nœuds.

Pour un nœud OU considérons d'abord le cas de la requête A OU B avec deux documents, l'un contenant chaque terme assez proches (cf. $d1$ sur la figure 1) et l'autre contenant deux occurrences de A côte à côte (cf. $d2$). Pour ce besoin d'information, utiliser A ou B dans le texte possède la même signification nous souhaitons donc obtenir la même fonction de proximité pour $d1$ et $d2$ avec la requête A OU B. En posant $p^d_{A\,OU\,B}(x) = \max(p^d_A(x), p^d_B(x))$ cette contrainte est vérifiée et nous généralisons ceci en posant $p_{q\,\mathrm{OU}\,q'} = \max(p_q, p_{q'})$ pour un nœud où les fils ne sont pas simplement des termes. Ceci correspond à l'opération faite dans le modèle flou. Par analogie, pour un opérateur ET, nous posons $p_{q\,\mathrm{ET}\,q'} = \min(p_q, p_{q'})$. La dernière étape qui consiste à déterminer le score de pertinence est détaillée dans la section 4.

3.2 Modèle à pertinence locale

Dans cette approche, nous considérons que les occurrences des termes apportent un élément de pertinence locale autour de leur position. Le signal de pertinence est aussi représenté par une fonction d'influence. Étant donné un terme et un document, pour accumuler les informations de pertinence à chaque position, nous additionnons les valeurs des pertinences locales calculées pour chaque occurrence du terme. La pertinence locale en une position x du texte est exprimée par $r^d_t(x) = \sum_{i \in Occ(t,d)} f(x-i)$. Chacune des feuilles de l'arbre de la requête porte un signal global qui représente la pertinence locale au terme pour chaque position du document. Au moment de l'évaluation de la requête nous devons donc combiner ces signaux en fonction du type d'opérateur (OU ou ET). Considérons d'abord le cas d'une requête disjonctive (cf. partie signal de la figure 1), quelque soit le terme à considérer nous souhaitons pour une telle requête prendre en compte le signal de tous les termes retrouvés afin qu'il contribue au calcul du score. Pour accumuler les informations de pertinence, nous posons pour l'opérateur OU $r_{q\,\mathrm{OU}\,q'} = r_q + r_{q'}$. Une fonction doit aussi être utilisée sur les nœuds ET. Si nous appliquons la fonction min comme dans le cas de la proximité floue, pour la requête A ET (B OU C) nous aboutissons à une incohérence. Par exemple, pour une position x dans un document, si nous avons les valeurs de pertinence locale égales à $r^d_A = 0.5$, $r^d_B = 0.4$ et $r^d_C = 0.8$, nous obtenons $r_{(A\,\mathrm{AND}\,(B\,\mathrm{OR}\,C))} = \min(0.5, 0.4 + 0.8) = 0.5$ et $r_{(A\,\mathrm{AND}\,B)OR(B\,\mathrm{AND}\,C))} = \min(0.5, 0.4) + \min(0.4, 0.8) = 0.8$. Par conséquent, les lois de Morgan ne sont pas respectées et nous ne pouvons pas utiliser cette fonction aux nœuds ET de l'arbre. Pour une requête conjonctive, en une position x du document, pour les raisons expliquées ci-dessus nous posons pour l'opérateur ET $r_{q\,\mathrm{ET}\,q'} = r_q \cdot r_{q'}$.

4 Détermination du score d'un document

Autant pour le modèle à proximité floue que pour celui à pertinence locale, l'évaluation d'une requête est effectuée en partant des feuilles. Tout d'abord, nous calculons pour chaque terme de la requête (pour les feuilles de l'arbre) la valeur de pertinence locale (resp. proximité floue) à chaque position x du document. Ensuite, nous évaluons ces valeurs au niveau de chaque nœud de l'arbre en appliquant (toujours pour chaque position x dans le document) les opérations correspondant aux deux opérateurs (ET ou OU). Finalement, en remontant jusqu'à la racine, nous obtenons le résultat, p^d_q (resp.

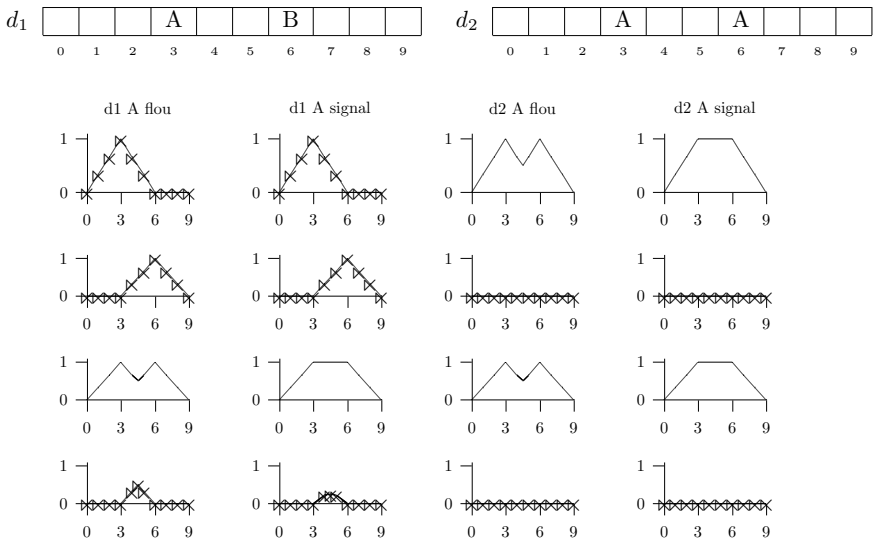

FIG. 1 – *Pour d*1 *et d*2 *– proximité floue et pertinence relative (signal).* 1ere *et* 2e *lignes termes A et B ;* 3e *et* 4e *lignes requêtes (A ou B) puis (A etB).*

r_q^d), qui permet de déterminer le score du document pour une requête.

La dernière étape après le calcul de p_q^d (resp. r_q^d) consiste à déterminer le score de pertinence $s(q, d)$ pour le document d par rapport à la requête q. Dans le cas du modèle booléen, le score, résultant de l'évaluation de la requête est binaire. Pour le modèle vectoriel, les formules de calcul de pertinence sont des produits scalaires ou des cosinus qui comportent une sommation qui peut s'interpréter comme une accumulation d'éléments de pertinence. Les méthodes du calcul intégral permettent de mettre en œuvre cette idée en calculant la surface en dessous d'une courbe, le score étant représenté par une courbe prenant les valeurs de proximité floue (resp. pertinence relative) à chaque position du document, nous l'exprimons ainsi : $s(q, d) = \int_{-\infty}^{+\infty} p_q^d(x)\ \mathrm{d}x$. Finalement, le score appartient à $\mathbb{R}^+$, ce qui permet de classer les documents par ordre décroissant de score, et dépend de l'influence de chaque occurrence, ce qui permet de prendre en compte la position relative entre les termes correspondant à la seconde idée de Luhn.

5 Conclusion

À partir de notre hypothèse : *les documents ayant des occurrences des termes de la requête proches doivent être classés en premier*, nous avons détaillé notre modèle à « zone d'influence » utilisant des requêtes booléennes. Par ailleurs, notre modèle offre l'avantage de prendre en compte les modèles classiques de recherche d'informations

comme le modèle booléen et le modèle vectoriel en contrôlant la valeur du paramètre k car ce dernier permet de régler la portée de l'influence des occurences de termes. Une valeur de l'ordre de 5 permet de spécifier une proximité de l'ordre de l'expression, une valeur de 15 à 30 la situe au niveau de la phrase et une valeur de l'ordre de 100 la porte au niveau du paragraphe. Prendre la limite lorsque $k \to +\infty$ permet de retrouver le modèle booléen classique. Prendre $k = \frac{1}{2}$ permet de retrouver le modèle du niveau de coordination précurseur du modèle vectoriel. Par conséquent, notre modèle met en œuvre notre objectif premier de donner un score en fonction de la localisation des termes de la requête dans les documents mais peut aussi être paramétré pour retrouver les comportements des méthodes traditionnelles de recherche d'informations.

Références

R. Baeza-Yates et B. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press / Addison-Wesley, 1999. ISBN 0-201-39829-X.

C. L. A. Clarke, G. V. Cormack, et E. A. Tudhope. Relevance ranking for one to three term queries. *Information Processing and Management*, 36 :291–311, 2000.

D. Hawking et P. Thistlewaite. Proximity operators - so near and yet so far. In D. K. Harman, editor, *TREC-4 proceedings*, 1995.

H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2 :159–168, 1958.

A. Mercier. Etude comparative de trois approches utilisant la proximité entre les termes de la requête pour le calcul des scores des documents. In *INFORSID*, 2004.

Patrick C. Mitchell. A note about the proximity operators in information retrieval. In *meeting on Programming languages and information retrieval*, pages 177–180. ACM Press, 1973.

Sadaaki Miyamoto. Fuzzy sets in information retrieval and cluster analysis. 1990.

Y. Rasolofo et J. Savoy. Term proximity scoring for keyword-based retrieval systems. In *ECIR 2003 proceedings*, pages 207–218, 2003.

G. Salton, E. A. Fox, et H. Wu. Extended Boolean information retrieval. *Communications of the ACM*, 26(11) :1022–1036, 1983.

Summary

Extracting, scoring and ranking documents relevant to a query is a main objective in the information retrieval domain. Our study focuses on the terms localization : the more the query terms occurrences are found close the more the document must be in the toplist. We present our model «area of influence» which scores the documents according to the terms localization. We detail the two alternatives, the first is based on a fuzzy proximity concept and the second on a local relevance concept.