Modèles de proximité : conception et comparaison à une méthode de recherche de passages

Annabelle Mercier, Amélie Imafouo, Michel Beigbeder

Modèle de proximité :

Conception et comparaison à une méthode de recherche de passages

Annabelle MERCIER, Amélie IMAFOUO, Michel BEIGBEDER
Categorie Chercheur.

Ecole Nationale Supérieure des Mines de Saint-Etienne
158, cours Fauriel
F 42023 SAINT ETIENNE CEDEX 2
{mercier,imafouo,mbeig}@emse.fr

RÉSUMÉ. *Notre étude se situe dans le domaine de la recherche d'informations. Certains modèles classiques comme le modèle vectoriel permettent de classer les documents par ordre de pertinence alors que d'autres, qui n'offrent pas cette possibilité de classement, possèdent des fonctionnalités particulières comme repérer les documents où les termes de la requête apparaissent proches les uns des autres. Cette dernière idée ayant conduit à des améliorations des résultats, nous formulons l'hypothèse que plus les occurrences des termes d'une requête se retrouvent proches dans un document, plus ce document doit être positionné en tête de la liste de réponses retournées par un système. Par ailleurs, les méthodes de recherche de passages permettent de sélectionner les passages répondant à la requête et définissent ainsi une proximité au niveau du passage. Après avoir rappelé les principaux modèles de recherche d'informations ainsi que les approches qui traitent de la proximité entre les termes nous présentons notre modèle de proximité qui utilise la logique floue. Nous avons expérimenté notre modèle et nous comparons les résultats obtenus par rapport à ceux d'une méthode de recherche de passages.*

ABSTRACT. *Scoring and ranking relevant documents compared to the user's information need is a principal objective in information retrieval domain. We base our study on the terms proximity over documents, we assume that more the occurrences of the query terms request are found close in a document then more this document must be positioned in the top of the document retrieve list. After a brief description of traditional models and the few approaches which treat proximity between the occurrences of terms, we present our method on fuzzy proximity which makes it possible to give a score of relevance to the documents according to the localization of the terms in these documents. In this study, we compare our approch with passage retrieval methods based on the* MG *and* LUCY *information retrieval system. We define a method to split the collection in passages.*

MOTS-CLÉS : *logique floue, proximités des termes, recherche de passages, recherche d'informations.*

KEYWORDS: *fuzzy logic, term proximity, passage retrieval, information retrieval.*

1. Introduction

Le domaine de la recherche d'informations depuis le début de l'informatisation des fonds documentaires jusqu'au développement actuel de la toile a vu apparaître de nombreuses techniques pour accéder à l'information. Une des premières idées, fondatrices du domaine et formulée par Luhn [LUH 58] en 1958, propose d'utiliser la fréquence des mots ainsi que leur position relative pour mesurer la pertinence d'un article par rapport à un besoin d'informations. A ce jour, beaucoup de modèles et de systèmes fondés sur l'idée de fréquence ont été développés, cependant, peu d'approches sont basées sur la proximité entre les termes pour détecter les documents pertinents.

Le modèle booléen est un modèle classique en recherche d'informations. Les documents y sont représentés par l'ensemble des termes qui le compose et les requêtes sont formulées à l'aide d'une expression booléenne qui est souvent représentée par un arbre. Pour une requête et un document donnés, une telle expression est évaluée soit à faux, soit à vrai. Par conséquent, le critère de décision de pertinence est binaire et les documents retournés par un système basé sur ce modèle ne peuvent pas être classés ce qui est un inconvénient majeur pour l'utilisateur. Un opérateur de proximité a été introduit dans une extension de ce modèle pour préciser une distance maximale entre deux termes de la requête retrouvés dans les documents. Cependant, cette extension a atteint aussi ses limites car d'une part, l'opérateur de proximité ne peut s'appliquer qu'aux feuilles de la requête (sa généralisation aux sous-arbres est inconsistante [MIT 73]), et d'autre part, la pertinence est toujours binaire.

Un autre modèle bien connu en recherche d'informations est le modèle vectoriel. Ce dernier prend en compte le poids des termes retrouvés dans le document pour l'attribution du score. Généralement, 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, c'est-à-dire du nombre de documents où le terme t apparaît. Les documents tout comme les requêtes sont représentés chacun par un vecteur, ce qui permet de calculer la valeur de similarité par la méthode du cosinus pris entre les deux vecteurs. En attribuant un score de similarité, ce modèle possède l'avantage de classer les documents, ces derniers sont présentés à l'utilisateur par ordre décroissant de pertinence calculée par le système.

Notre étude propose un nouveau modèle de recherche d'informations qui utilise la logique floue dans sa conception et qui attribue un score de pertinence aux documents en fonction de la position relative des termes de la requête. Afin de mettre en évidence notre utilisation de la proximité, nous avons réalisé une série d'expériences pour comparer notre modèle à un méthode de recherche de passage qui dans une certaine mesure se rapproche de l'utilisation de la proximité.

Après un état de l'art sur les méthodes de proximité et de recherche de passages, nous présentons notre modèle puis nos expériences et enfin, nous terminons par quelques perspectives.

2. État de l'art

Habituellement, un système de recherche d'informations est basé sur un modèle. Il utilise une méthode d'indexation qui s'appuie sur les occurrences des termes trouvés dans les documents pour construire une représentation des documents. Dans la suite, nous appelons T l'ensemble des termes et D l'ensemble des documents.

2.1. *Modèles classiques*

Nous avons déjà rapidement présenté en introduction les modèles booléen et vectoriel, classiques en recherche d'informations. Les limites du modèle booléen ont conduit à la conception de modèles booléens étendus pour présenter les documents par ordre décroissant de pertinence calculée par le système. Par ailleurs, plusieurs modèles fondés sur la théorie des sous ensembles flous ont été développés [MIY 90] et permettent de graduer le score par rapport au modèle booléen de base en attribuant un score de pertinence normalisé. Dans ces modèles, à 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 exprimée à l'aide d'une expression booléenne qui est représentée par un arbre. Les nœuds portent les opérateurs ET et OU et les feuilles les termes. Pour l'appariement entre un document et une requête, un nœud avec l'opérateur ET (resp. OU) est évalué en prenant le minimum (resp. maximum) 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.

2.2. *Méthodes à intervalles*

D'autres approches utilisent directement la proximité des termes pour le calcul du score des documents [CLA 00, HAW 95, RAS 03]. Ces méthodes procèdent d'abord par une phase de sélection des intervalles du texte qui contiennent les termes de la requête. Les intervalles sélectionnés diffèrent selon les méthodes. La méthode de Clarke et al. [CLA 00, CLA 95, CLA 96] sélectionne les intervalles les plus courts qui contiennent tous les termes de la requête. Si le nombre de documents contenant tous les termes n'est pas suffisant, cette contrainte est relaxée, les documents ayant moins de termes sont retenus jusqu'à ce que le seuil du nombre de documents à retrouver soit atteint. Cette méthode prend en compte moins d'intervalles que les suivants et les intervalles sélectionnés ne sont pas emboîtés les uns dans les autres. Dans la méthode de Hawking et al. [HAW 95], pour chaque occurrence d'un terme de la requête, l'intervalle le plus court débutant sur cette occurrence et contenant tous les autres termes est sélectionné. Enfin Rasolofo et al. [RAS 03] choisissent de sélectionner les intervalles contenant deux termes de la requête, avec la contrainte supplémentaire que leur longueur soit de moins de cinq mots. Un score est ensuite attribué à tous ces intervalles

(plus l'intervalle est court, plus son score est élevé), et enfin le score d'un document dépend de la somme des scores des intervalles sélectionnés. Une étude préliminaire a montré que les résultats obtenus avec ces méthodes sont meilleurs que ceux des modèles traditionnels de recherche d'informations et nous avons déjà comparé notre approche aux modèles à base d'intervalles [MER 04].

2.3. *Recherche de passages*

L'augmentation du nombre de documents accessibles électroniquement a conduit à l'émergence des méthodes de recherche de passages car les méthodes de recherche traditionnelles n'était pas adaptées au traitement des documents en texte intégral mais plutôt aux résumés d'articles. Par exemple, le modèle vectoriel et sa technique de pondération permet de rechercher efficacement à l'aide de requêtes courtes, car pour une idée générale il est plus facile de retrouver les mot-clés dans les résumés.

Nous distingons trois type de méthodes de passages, ces dernières s'appuient sur la structure logique et utilisent le balisage du texte [JUS 95, SAL 93], sur le vocabulaire employé [HEA 97, HEA 93], ou sur un découpage en fenêtre [CAL 94, KAS 97].

Les méthodes de passages comme celle de Hearst se préoccupent des idées dites «secondaires» du texte en sélectionnant les morceaux de texte pertinents qui contiennent les termes de la requête avec une fréquence localement élevée. Connue sous le nom de «TextTiling» [HEA 97, HEA 93], cette méthode appartient à la famille de méthodes de recherche de passages basées sur la sémantique du texte. Il existe d'autres méthodes qui se fondent sur la structure du texte ou bien qui utilisent les balises spécifiques des documents textuels structurés afin d'en effectuer le découpage en passages. Ces méthodes s'appuient sur les pages ou les sections [JUS 95], sur les paragraphes [SAL 93] ou sur des fenêtres variables ou non de mots [CAL 94]. Kaszkiel et Zobel [KAS 97] utilisent des passages de longueur fixe pour classer les documents et montrent que leur méthode est aussi efficace qu'une méthode vectorielle traditionnelle [BAE 99]. D'une part, la méthode que nous expérimentons (cf. section 4) fait partie de cette dernière famille car les passages sont constitués d'un nombre fixe de lignes, et d'autre part, nous nous inspirons des mesures données par Wilkinson [WIL 94] pour calculer le score d'un passage.

En général, une méthode de recherche de passages permet de sélectionner les informations à un niveau de granularité inférieur à celui du document ce qui permet aussi d'accéder aux idées secondaires cachées dans un texte intégral. Avec ce genre de méthode, il serait intéressant de permettre à l'utilisateur de distinguer les sujets principaux des sujets secondaires, ainsi, il formulerait une requête à deux niveaux où les sujets secondaires seraient situés dans le contexte des sujets principaux. Par conséquent, dans une requête conjonctive, on pourrait distinguer les termes qui doivent être retrouvés ensembles de ceux qui doivent apparaître dans le contexte de l'autre. Justement, nous pensons que l'utilisation de la proximité dans le calcul de la pertinence permet de détecter ces cas en prenant en compte la position relative entre les mots.

Dans les méthodes de recherche de passages, le découpage du texte intégral fournit une proximité au niveau du passage, c'est pourquoi nous avons voulu les comparer à notre méthode.

3. Modèle basé sur la proximité

3.1. *Zone d'influence d'un terme*

Les modèles booléen et vectoriel procèdent avec une approche *globale* de l'influence des occurrences d'un terme pour mesurer la pertinence d'un document à une requête, en considérant, pour le premier modèle, l'appartenance du terme au document et en prenant en compte dans le second, la fréquence des termes. Dans ces modèles, la position relative des termes de la requête retrouvés dans les documents n'est donc pas prise en compte pour le calcul de pertinence. Cependant, le sens d'un texte ne dépend pas seulement du vocabulaire employé mais aussi de la proximité spatiale de ses termes. Par conséquent, nous adoptons une approche *locale* dans le sens où nous modélisons une *influence* des occurrences que nous appelons **proximité floue au terme**. Cette proximité est graduée et nous permet de savoir si en un endroit du texte, on est plus ou moins proche d'une occurrence d'un terme de la requête.

L'influence d'un mot est représentée par 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 $\mathbb{R}^+$. Nous pouvons utiliser différentes sortes de fonctions d'influence comme les fonctions de Hamming, les fonctions gaussiennes ou bien les fonctions triangulaires. De plus, pour chacune de ces fonctions, nous pouvons régler un certain nombre de paramètres afin d'obtenir une fonction différente associée à chaque terme de la requête. Dans les expériences réalisées, nous utilisons une fonction triangulaire, et nous introduisons un paramètre k pour contrôler la largeur de la zone d'influence des termes ainsi, la valeur en un point x du texte est égale à 1 puis décroit de $\frac{1}{k}$ aux positions voisines jusqu'à atteindre la valeur 0. Pour une occurrence d'un terme, la translation $g(x) = f(x - i)$ d'une fonction d'influence f permet de modéliser la proximité floue. Une fonction d'influence est exprimée par :

$$f(x) = \max(\frac{k - |x|}{k}, 0).$$

Les requêtes sont des expressions booléennes comme dans le modèle utilisant les sous-ensembles flous présenté dans la section 2.1. Leur évaluation est effectuée en partant des feuilles. Tout d'abord, pour chaque position x du document, la proximité floue est calculée pour chaque terme de la requête c'est-à-dire pour les feuilles de l'arbre. Ensuite, ces valeurs sont aussi évaluées au niveau de chaque nœud de l'arbre en appliquant (toujours pour chaque position x dans le document) les fonctions correspondant aux deux opérations (ET ou OU). Finalement, étant donnée la requête (c'est-à-dire l'arbre), le score du document est déterminé en fonction du résultat obtenu à la racine.

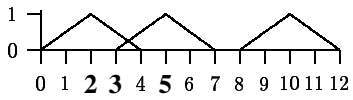

Figure 1. *Représentation de la proximité floue d'un terme. Par exemple, à la position $x = 3$, la valeur de proximité floue retenue est celle de la plus proche occurrence, par conséquent, il s'agit de la valeur maximale entre les valeurs de proximité floue des occurences aux positions $x = 2$ et $x = 5$ soit celle correspondant à l'occurrence du terme en $x = 2$.*

3.2. *Proximité floue d'un terme*

La proximité floue à un terme t en une position x d'un document est déterminée par les occurrences de ce terme qui se situent aux positions environnantes. Par conséquent, la valeur de proximité floue à un terme t en une position x d'un document sera la valeur de la proximité de la plus proche occurrence du terme t.

Par exemple, considérons la position $x = 3$ de la figure 1, nous attribuons comme valeur de proximité floue celle provenant de la fonction d'influence de l'occurrence du terme le plus proche. Les fonctions d'influence étant décroissantes par rapport à la distance des occurrences, en une position x du texte la valeur de proximité floue maximale est prise :

$$p_t^d(x) = \max_{i \in Occ(t,d)} f(x - i)$$

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 aux termes correspondants. 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 (cf. 1^{ere} et 2^e courbes de la figure 2).

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 les deux termes une fois côte à côte (cf. fig. 2) et l'autre contenant deux occurrences de A côte à côte (cf. fig. 3).

Pour ce besoin d'information, l'utilisation de A ou de B dans le texte par un auteur a la même signification, nous souhaitons donc obtenir la même fonction de proximité pour ces deux documents (comme le montre la troisième courbe des figures 2 et 3). 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

$$p_{q\,\mathrm{OU}\,q'} = \max(p_q, p_{q'}).$$

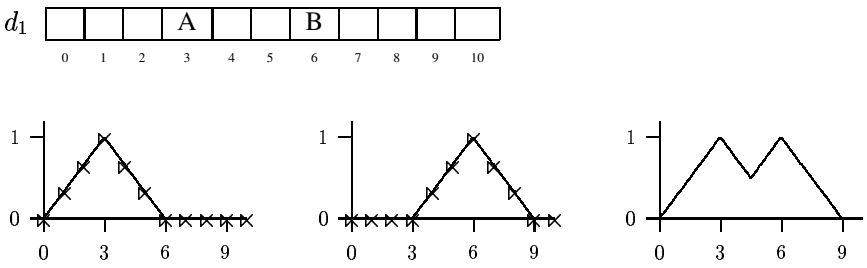

Figure 2. *Document 1 – la première courbe montre la représentation de la proximité floue pour le terme A et la seconde pour le terme B, la troisième courbe correspond à la proximité floue calculé pour la requête* A OU B.

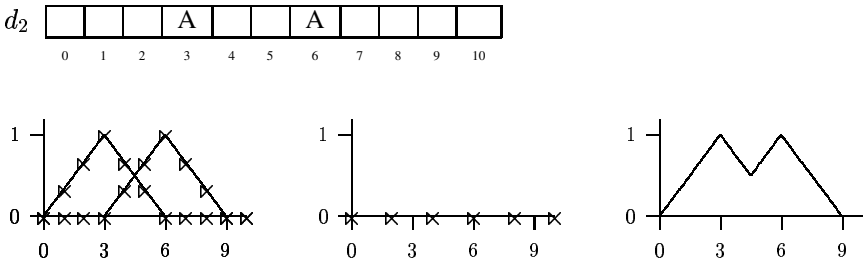

Figure 3. *Document 2 – la* 1^{re} *(resp.* 2^e*) courbe montre les valeurs de proximité floue pour chaque occurrence du terme A (resp. B), la* 2^e *courbe (resp.* 3^e*) montre la proximité floue calculée pour le terme B (resp. A), la* 3^e *courbe correspond aussi à la proximité floue calculée pour la requête* A OU B.

pour un nœud où les fils ne sont pas simplement des termes. Ceci correspond bien à l'opération faite dans le modèle flou (cf. section 2.1). Par analogie, pour un opérateur ET, nous posons

$$p_{q\,\mathrm{ET}\,q'} = \min(p_q, p_{q'}).$$

De plus, notre modèle peut s'adapter facilement en employant les autres fonctions appliquées en logique floue pour les opérateurs ET et OU [BOU 95, ZAD 78]. Par exemple, l'opération à effectuer pour un OU peut être $max(x + y - 1, 0)$, dans ce cas, l'opération associé à l'opérateur ET est $\min(x + y, 1)$ avec x et y appartenant à deux sous-ensemble flous différents.

3.3. *Score d'un document*

La dernière étape après le calcul de p_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 est binaire, c'est le résultat de l'évaluation de la requête pour le document d. Dans le cas du 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 de calcul intégral permettent de mettre en œuvre cette idée d'accumulation de pertinence. En considérant que nous représentons le score par une courbe prenant les valeurs de proximité floue à chaque position du document, nous calculons le score comme étant la surface en dessous de cette courbe

$$s(q, d) = \int_{-\infty}^{+\infty} p_q^d(x)\ \mathrm{d}x.$$

Dans nos expériences, nous avons discrétisé la proximité floue et concrètement pour le calcul du score, nous additionnons la valeur de proximité floue à chaque position x du document.

Les scores obtenus sont positifs, ce qui permet de classer les documents par ordre décroissant de pertinence. Nous mettons donc en œuvre la seconde idée de Luhn car le score obtenu dépend de la position relative en calculant la proximité floue de chaque terme.

4. Expérimentations

4.1. *Collection de test utilisée*

Pour notre expérimentation, nous utilisons la collection de la campagne d'évaluation de la conférence TREC [1].

En particulier, nous utilisons la collection de test WT10g[2] avec les besoins d'informations 451–500[3] et les jugements de pertinence [4] correspondant.

Les requêtes sont construites avec les mots-clefs du champ «TITRE». Pour chaque topiques, ces requêtes plates sont utilisées pour obtenir les résultats pour le modèle vectoriel avec l'outil MG [WIT 99] et pour la mesure Okapi BM-25 [ROB 92] avec LUCY [5]. Pour les méthodes de proximité, nous effectuons une première sélection des documents de la collection en appliquant la requête disjonctive correspondante, ensuite, les documents sont de nouveau classés en fonction des scores obtenus pour cha-

1. http://trec.nist.gov/
2. http://es.csiro.au/TRECWeb/wt10g.html
3. http://trec.nist.gov/data/topics_eng/topics.451-500.gz
4. http://trec.nist.gov/data/qrels_eng/qrels.trec9.main_web.gz
5. /http://www.seg.rmit.edu.au/lucy

cune des méthodes (méthode de Clarke, Hawking et Rasolofo). Pour notre méthode de proximité floue, nous utilisons aussi ces requêtes disjonctives.

La collection que nous avons utilisée a été construite à partir de la collection WT10G. Sur cette nouvelle collection, les documents pertinents sont répartis de façon uniforme : ainsi, en sélectionnant n'importe quelle portion de collection, la même distribution de documents pertinents soit obtenue et le nombre de documents pertinents par besoin d'informations (et le nombre de documents pertinents pour tous les besoins d'information) est proportionnel à la taille de la portion. L'intérêt d'une telle collection est qu'il est possible de la découper en portions (sous-collections) de taille croissante, d'étudier des propriétés de recherche d'informations sur chaque portion, par exemple les métriques d'evaluation, et d'analyser l'influence de la taille de la portion sur ces propriétés. Une telle étude fait partie des perspectives de ce travail. La méthodologie générale qui soutend la construction de cette collection a été décrite dans [IMA 05]. Les résultats présentés dans la section suivante ont été obtenus sur la sous-collection (WT10G_1400) de WT10G uniforme qui contient 1400000 documents.

4.2. *Simulation d'une méthode de recherche de passages*

Nous avons découpé les documents en passages de N lignes chacun. Dans notre expérimentation, le nombre de lignes constituant un passage est $N = 45$ et la valeur de la constante k qui définie la portée de la proximité pour notre méthode est 50. Nous avons ensuite indexé et interrogé cette collection de passages en utilisant l'outil MG basé sur le modèle vectoriel (resp. l'outil LUCY utilisant la méthode Okapi) , en considérant chaque passage comme un document. Pour chaque topique, nous obtenons un score pour chaque passage. Pour calculer le score d'un document, nous combinons les scores des différents passages de ce document, nous nous sommes inspirées des mesures proposées par Wilkinson [WIL 94]. Nous avons utilisé plusieurs fonctions pour combiner le score des passages afin de calculer le score du document :

– le $maximum$ des scores des passages d'un document,

– la $somme$ des scores des passages d'un document,

– le nombre ($compte$) de passages du document ayant un score positif,

– la $moyenne$ des scores des passages d'un document.

4.3. *Résultats*

La figure 5 montre les courbes rappel/précision pour les méthodes de passages basées sur MG, la méthode de proximité floue et la méthode vectorielle de base. Pour tous les niveaux de rappel, la méthode de proximité floue est la plus performante en terme de précision. La méthode vectorielle de base procure en moyenne de meilleurs résultats que les méthodes de passages. Pour ces dernières, lesméthodes utilisant la $somme$ et le $compte$ pour le calcul du score sont plus performantes que les autres.

La figure 6 montre les courbes rappel/précision pour les méthodes de passages basées sur LUCY, la méthode de proximité floue et la méthode Okapi classique. Aux deux premiers niveaux de rappel, la méthode de proximité est au-dessus de la méthode Okapi. Pour les niveaux de rappel de 20% à 50%, la méthode Okapi est meilleure mais néanmoins la méthode de proximité reste très proche d'elle et passe à nouveau devant la méthode Okapi à partir de 60% de rappel. Les méthodes de passages basée sur LUCY possèdent le même comportement que les méthodes basée sur MG.

Pour comparer les résultats, nous avons seulement retenu les topiques pour lesquels toutes les méthodes ont retourné des documents, de ce fait, le tableau de la figure 4 se rapporte à 42 requêtes sur les 50 totales. Présenter les résultats de cette manière permet d'avoir une vision plus précise qui prend en compte les topiques individuellement sans pour autant présenter la courbe rappel/précision pour chaque topique. Le tableau montre l'ordre des méthodes aux trois premiers niveaux de rappel. Cet ordre est défini par rapport au taux de précision pour chacune des méthodes à un niveau de rappel. Nous présentons pour chaque méthode le nombre de topiques à un rang donné (1, 2 ou 3). Par exemple, la méthode de proximité floue est la meilleure pour 19 topiques sur 42 à 10% de rappel. Pour le premier rang, nous remarquons que la méthode de proximité floue est la plus efficace pour 35% à 45% des topiques, elle est suivie par la méthode Okapi qui est performante pour 26% à 35% des requêtes. Pour les méthodes de passages pour MG d'une part et pour LUCY d'autre part, nous avons regroupé les résultats obtenus avec les quatre fonctions ($maximum$, $somme$, $compte$, $moyenne$). Pour quelques topiques, la méthode de passages donnent de bons résultats 11% au rang 1, 14% au rang 2 et 38% au rang 3, cela signifie que cette méthode passe devant les méthodes usuelles pour certains topiques. Par conséquent, l'usage direct de la proximité apportent des améliorations significatives.

5. Conclusion

Dans notre étude, nous avons tout d'abord présenté les modèles classiques tout en faisant le lien avec l'utilisation de la proximité des occurrences des termes en recherche d'information. Ensuite, 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 méthode de proximité floue utilisant des requêtes booléennes. Notre méthode définit une proximité aux alentours de quelques mots alors que les méthodes de passages définissent une proximité au niveau du passage. Nous avons vu que les résultats obtenus pour la méthode de proximité floue sont meilleurs que ceux des autres modèles expérimentés aux premiers niveaux de rappel. Néanmoins, les méthodes à bases de passages donnent de bons résultats pour certaines requêtes. Par conséquent, l'utilisation de la proximité apporte une amélioration pour les documents retournés en début par un système de recherche d'information. Notre expérimentation rejoint les conclusions d'une étude antérieure où nous avons remarqué que les méthodes basées sur les intervalles (méthode de Clarke, de Hawking et de Raso-

rang	proximite floue	vectoriel	okapi	Passage LUCY	Passage MG
Niveau 0					
1	15	9	11	2	5
2	10	7	14	6	5
3	6	12	5	10	9
Niveau 10 % de rappel					
rang	proximite floue	vectoriel	okapi	Passage LUCY	Passage MG
1	19	6	13	3	1
2	10	10	15	3	4
3	6	15	5	6	10
Niveau 20 % de rappel					
rang	proximite floue	vectoriel	okapi	Passage LUCY	Passage MG
1	19	5	15		3
2	7	9	16	4	6
3	4	17	3	2	16

Figure 4. *Ordre des méthodes aux trois premiers niveaux de rappel pour 42 requêtes. Nous comptabilisons par méthode le nombre de topiques pour les lesquels la méthode se situe au premier, second et troisième rang. Le premier rang correspond au plus fort taux de précision, graphiquement cela revient à regarder quelle est la première courbe cést-à-dire celle ayant la plus grande valeur de précision à un niveau de rappel donné, nous faisons de même pour les rangs 2 et 3.*

lofo décrites dans la section 2.2) obtiennent de meilleurs résultats que les méthodes classiques aux premiers niveaux de rappel.

6. Bibliographie

[BAE 99] BAEZA-YATES R., RIBEIRO-NETO B., *Modern Information Retrieval*, ACM Press / Addison-Wesley, 1999.

[BOU 95] BOUCHON-MEUNIER B., *La logique floue et ses applications*, Addison-Wesley, 1995.

[CAL 94] CALLAN J. P., « Passage-level evidence in document retrieval », *SIGIR 94 proceedings*, Springer-Verlag New York, Inc., 1994, p. 302–310.

[CLA 95] CLARKE C. L. A., CORMACK G. V., BURKOWSKI F. J., « Shortest Substring Ranking », *The Fourth Text REtrieval Conference (TREC-4)*, 1995.

[CLA 96] CLARKE C., CORMAK G., « Interactive Substring Retrieval : MultiText Experiments for TREC5 », *The Fifth Text REtrieval Conference*, 1996.

[CLA 00] CLARKE C. L. A., CORMACK G. V., TUDHOPE E. A., « Relevance ranking for one to three term queries », *Information Processing and Management*, vol. 36, 2000, p. 291-311.

[HAW 95] HAWKING D., THISTLEWAITE P., « Proximity Operators - So Near And Yet So Far », HARMAN D. K., Ed., *TREC-4 proceedings*, 1995.

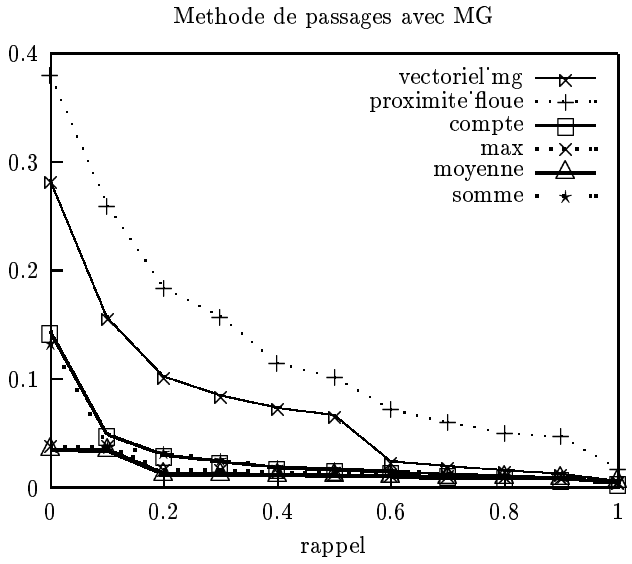

Figure 5. *Evaluation de la collection découpée en passages avec* MG *: modèle vectoriel, proximité floue et méthodes de passages*

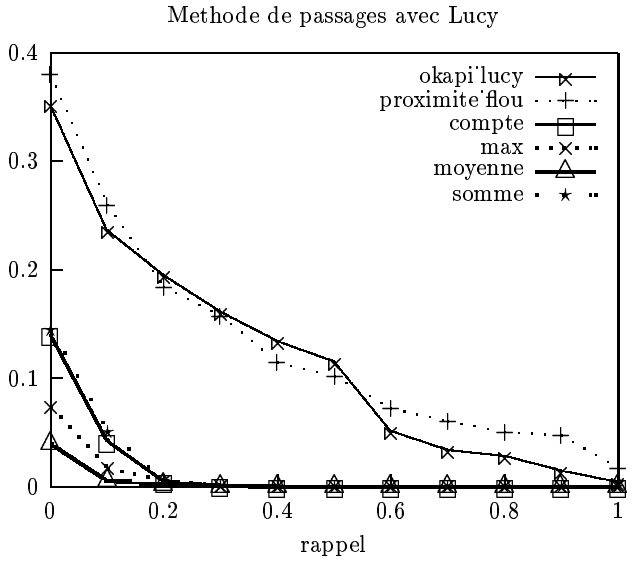

Figure 6. *Evaluation de la collection découpée en de passages avec* LUCY *: okapi lucy, proximité floue et méthodes de passages*

[HEA 93] HEARST M. A., PLAUNT C., « Subtopic structuring for full-length document access », *SIGIR 93 proceedings*, ACM Press, 1993, p. 59–68.

[HEA 97] HEARST M. A., « TextTiling : segmenting text into multi-paragraph subtopic passages », *Comput. Linguist.*, vol. 23, n° 1, 1997, p. 33–64, MIT Press.

[IMA 05] IMAFOUO A., BEIGBEDER M., « Passage à l'échelle :une méthodologie pour l'étude de l'influence du volume de collection sur les modèles de RI », *Actes de la 2ème conférence Francophone en Recherche d'Information-CORIA'05*, Mars 2005.

[JUS 95] JUSTIN ZOBEL ALISTAIR MOFFAT R. W., SACKS-DAVIS R., « Efficient retrieval of partial documents », , !995, p. 361-377.

[KAS 97] KASZKIEL M., ZOBEL J., « Passage Retrieval Revisted », *SIGIR 97 proceedings*, 1997.

[LUH 58] LUHN H. P., « The automatic creation of literature abstracts », *IBM Journal of Research and Development*, vol. 2, 1958, p. 159-168.

[MER 04] MERCIER A., BEIGBEDER M., « Experiments with different proximity based information retrieval models », *2004 International Conference on Advances in Intelligent Systems proceedings*, November 2004.

[MIT 73] MITCHELL P. C., « A note about the proximity operators in information retrieval », *meeting on Programming languages and information retrieval*, ACM Press, 1973, p. 177–180.

[MIY 90] MIYAMOTO S., « Fuzzy Sets in information retrieval and cluster analysis », , 1990, KluwerAcademic Publishers.

[RAS 03] RASOLOFO Y., SAVOY J., « Term Proximity Scoring for Keyword-based Retrieval Systems », *ECIR 2003 proceedings*, 2003, p. 207–218.

[ROB 92] ROBERTSON S. E., WALKER S., HANCOCK-BEAULIEU M., GULL A., LAU M., « Okapi at TREC », *Text REtrieval Conference*, 1992, p. 21-30.

[SAL 93] SALTON G., ALLAN J., BUCKLEY C., « Approaches to passage retrieval in full text information systems », 1993, p. 49–58.

[WIL 94] WILKINSON R., « Effective retrieval of structured documents », *SIGIR 94 proceedings*, Springer-Verlag New York, 1994, p. 311–317.

[WIT 99] WITTEN I. H., MOFFAT A., BELL T. C., *Managing Gigabytes : Compressing and Indexing Documents and Images*, Morgan Kaufmann, 1999.

[ZAD 78] ZADEH L., « Fuzzy sets as a basis for a theory of possibility », *International Journal Fuzzy Sets Systems*, vol. 1, 1978, p. 3–28.