Chapitre IX. Organisation et accès à l'information pertinente *

1. Introduction

La requête initiale exprimée par l'utilisateur peut ne pas trouver de documents pertinents (ou peu de documents pertinents). Une méthode qui peut corriger ceci est la technique de retour de pertinence. Cette technique consiste à extraire l'information à partir des documents jugés (pertinents ou bien non pertinents) par l'utilisateur, pour modifier la requête initiale, afin d'améliorer la qualité des résultats.

D'un autre coté, la plupart des systèmes de recherche d'information renvoient en réponse à la requête de l'utilisateur une liste ordonnée de documents. Cette liste étant souvent très longue, les utilisateurs ne peuvent pas examiner tous les documents proposés. Cependant, de nombreux travaux ont montré qu'un découpage thématique des documents peut s'avérer plus efficace puisque cela permet d'individualiser les passages pertinents susceptibles d'être noyés dans une masse d'informations sans rapport avec le besoin exprimé par l'utilisateur.

Nous étudions ici une approche qui consiste à s'appuyer sur un processus de classification pour organiser la liste des réponses et d'améliorer la pertinence de la recherche. Afin d'obtenir des classes qui représentent au mieux les différents thèmes abordés par les documents, une nouvelle recherche est effectuée sur des zones documentaires plus fines que les documents ce qui permettra d'isoler l'information pertinente à l'intérieur d'un texte. Les résultats indiquent une amélioration sensible des résultats de la recherche d'information.

La section 2 détaille les différents éléments de nos approches pour l'organisation de l'information pertinente. Puisque le but est de réorganiser les résultats retournés par un SRI, cette section commence par présenter les résultats d'évaluation de la technique de retour de pertinence sur l'amélioration des performances. Dans la troisième section, l'évaluation effectuée concerne la capacité des méthodes de segmentation à mettre en valeur certains documents à partir de leur segmentation. Dans la quatrième section, nous explorons les différentes pistes pour considérer ces segments lors de la production de groupes thématiques de résultats. Nous discutons la manière de déduire des catégories thématiques de documents à partir de groupes de segments trouvés.

2. Retour de pertinence

Il est naturel, lorsqu'on l'on pose une requête à un système de recherche documentaire, d'enlever les termes de la requête ne paraissant pas donner de bons résultats soit parce qu'ils ne donnent pas de documents, soit par ce qu'ils en donnent trop. Le "retour de pertinence" est le nom donné à la méthode de modification automatique de la requête permettant cette fonctionnalité. Bien qu'on ait été tenté de mettre en place une technique de retour de pertinence sur la méthode booléenne, la recherche par similitude a vite constitué un domaine de recherche de prédilection pour cette technique.

Une fois sa requête effectuée, l'utilisateur peut sélectionner quelques textes qu'il considère comme pertinent. Dans un système de retour de pertinence automatique, le système affecte des poids à chaque terme de la requête en fonction de leur importance dans les documents

* Une grande partie de ce chapitre a été publiée dans les articles suivants :

1. F.Harrag, A. Hamdi-Cherif, A. S. Al-Salamn et E. El-Qawameh, Evaluating the Effectiveness of VSM model and Topic segmentation in Retrieving Arabic Documents, *International Journal of Computer Systems Science and Engineering*, ISSN 0267-6192, Vol.26, N.1, pp.55-68, January 2011.

sélectionnés. En plus de modifier les poids des différents termes, le système peut aussi modifier la requête initiale en supprimant ou en ajoutant des termes.

2.1. Probabilités de pertinence

Il est raisonnable de penser qu'un terme qui est présent fréquemment dans des documents jugés pertinents et très peu présent dans les documents jugés non pertinents soit jugé comme un bon terme pour une requête. Cette remarque peut permettre d'augmenter le poids de certains termes dans certaines requêtes. Elle a été étudiée au début des années 1970 par [Robertson et Sparck Jones, 1976].

Leurs travaux ont débouché sur une théorie probabiliste utilisant pour le calcul de poids le résultat suivant :

$$w_i = \log \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad (9.1)$$

Avec p_i probabilité que le terme t_i soit présent pour un document pertinent et q_i la probabilité que le terme t_i soit présent dans un document non pertinent. L'évaluation des probabilités n'étant pas très aisée, les probabilités sont habituellement remplacées par des calculs de fréquence d'apparition des termes dans le document ou l'ensemble documentaire.

De ce fait, si l'on considère que le terme t_i apparaît n_i fois dans les N documents de la collection et qu'il apparaît r_i fois dans les R (par défaut $R = 5$) documents pertinents de cette même collection, le calcul du poids de pertinence w_i se définit alors par :

$$w_i \approx \log \frac{r_i(N-n_i-R+r_i)}{(R-r_i)(n_i-r_i)} \qquad (9.2)$$

Tous les termes trouvés dans les R documents sont classés dans l'ordre décroissant par poids de pertinence w_i. Les poids des K premiers termes sont calculés et sont alors fusionnés avec les termes de la requête initiale pour créer une nouvelle requête. Quelques termes parmi les termes sélectionnés peuvent être dans la requête initiale. Pour les premiers termes sélectionnés qui ne sont pas dans la requête initiale, le poids est mis à 0.5. Pour ceux qui ont été déjà dans la requête initiale, le poids est mis à $0.5 * TF_t$, où TF_t est la fréquence d'apparition du terme t_i dans la requête initiale. Les termes sélectionnés sont fusionnés avec la requête initiale pour formuler une requête enrichie. Quand un terme sélectionné est l'un des termes de la requête initiale, son poids dans la requête enrichie est la somme de son poids dans la requête initiale et son poids assigné lors du processus de sélection des termes. Pour un terme sélectionné qui n'a pas été dans la requête initiale, son poids dans la requête finale est le même que son poids de sélection qui est égal à 0.5. Les poids des termes de la requête initiale qui n'ont pas été sélectionnés restent inchangés.

2.2. Calcul des poids de pertinence

Pour l'évaluation de l'impact de la technique de retour de pertinence sur les performances de notre système de recherche d'information, nous utilisons le même jeu de requêtes utilisé dans le chapitre 5. Les termes extraits des hadiths pertinents pour la requête R_3 (H_{5834}) sont utilisés pour le calcul des poids de pertinence. En fixons la valeur de K à 3, les termes sélectionnés sont : "سما", "كنا" et "قسم", leurs poids de pertinence, de sélection et d'expansion sont détaillés dans le tableau 9.1 ci-dessous. Le terme "قسم" a été ajouté à la requête initiale est son poids a été mis à 0.5, les poids des deux autres termes "سما", "كنا" existant initialement dans la requête R_3 ont été modifiés.

Terme de l'index	Poids de pertinence
قسم	6.70
سمّى	6.67
كنّى	4.69

Terme de l'index	Poids de sélection
قسم	0.50
سمّى	1.00
كنّى	1.00

Terme de l'index	Poids d'expansion
قسم	0.50
سمّى	10.00
كنّى	19.04

Tableau 9.1. Calcul des poids de pertinence, de sélection et d'expansion pour la requête R_3 (H_{5834}).

2.3. Calcul de similarité pour la requête après expansion

Le tableau 9.2 ci-dessous représente la liste des hadiths rapportés par le système de recherche d'information après expansion (enrichissement) de la requête R_3. L'ensemble des réponses a été enrichie par de nouveaux hadiths, ce même ensemble peut être divisé en deux classe de hadiths : pertinents et non pertinents.

Kitab	Bab	Num Hadith	Matn Hadith
العلم	إثم من كذب على النبي صلى الله عليه وسلم	110	**تسموا باسمي ولا تكتنوا بكنيتي**، ومن رآني في المنام فقد رآني حقا، ....
الخمس	باب: قول الله تعالى: {فأن لله خمسه وللرسول} /الأنفال: 41/.	2947	أحسنت الأنصار، **سموا باسمي** ولا **تكنوا بكنيتي**، فإنما أنا **قاسم**
الأدب	أحب الأسماء إلى الله عز وجل	5832	**سم** ابنك عبد الرحمن
الأدب	اسم الحزن	5836	أن أباه جاء إلى النبي صلى الله عليه وسلم فقال: (ما **اسمك**). قال: حزن، قال: (أنت سهل). قال: لا أغير **اسماً سمانيه** أبي.
الأدب	تحويل الاسم إلى اسم أحسن منه	5838	... قال: (ما **اسمه**). قال: فلان، قال: (ولكن **اسمه** المنذر). **فسماه** يومئذ المنذر
الأدب	أبغض الأسماء إلى الله	5852	أخنى **الأسماء** يوم القيامة عند الله رجل **تسمى** ملك الأملاك
صفة الصلاة	وجوب القراءة للإمام والمأموم في الصلوات كلها، في الحضر والسفر، وما يجهر فيها وما يخافت	722	...، يقال له أسامة بن قتادة، **يكنى** أبا سعدة، قال: أما إذ نشدتنا، فإن سعدا كان لا يسير بالسرية، ولا **يقسم** بالسوية، ولا يعدل في القضية ...
البيوع	ما قيل في اللحام والجزار	1975	...جاء رجل من الأنصار، **يكنى** أبا شعيب، فقال لغلام له قصاب...
أبواب العمل في الصلاة	يفكر الرجل الشيء في الصلاة	1163	ذكرت وأنا في الصلاة تبرا عندنا، فكرهت أن يمسي، أو يبيت عندنا، فأمرت **بقسمته**

Tableau 9.2. Liste des réponses pour la requête R_3 (H_{5834}) après expansion.

Le tableau 9.3 indique les résultats d'évaluation obtenus sur notre corpus prophétique selon la R_3 requête après expansion.

Doc	Pertinence	Précision	Rappel
3344.txt	1	1	0.14
3345.txt	1	1	0.28
3346.txt	1	1	0.42
5833.txt	1	1	0.57
2014.txt		0.8	0.57
2015.txt		0.66	0.57
110.txt	1	0.71	0.71
2947.txt	1	0.75	0.85
5832.txt	1	0.77	1

Tableau 9.3. Résultats d'évaluation pour la requête R_3 après expansion.

Les Figures 9.1 et 9.2 représentent la comparaison des niveaux de Rappel et de Précision pour la requête R_3 avec et sans expansion.

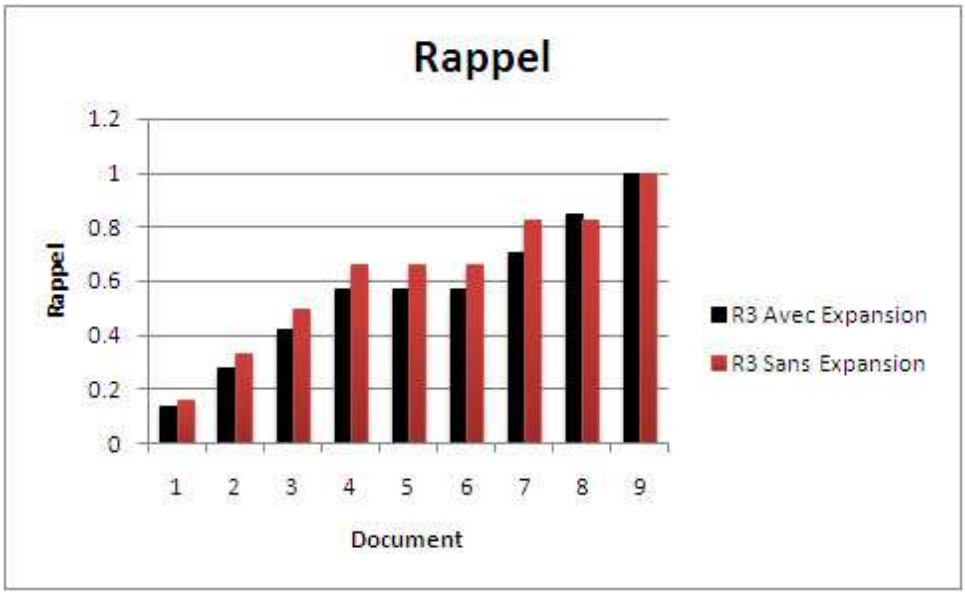

Fig. 9.1. Niveau de Rappel pour la requête R_3 avec et sans expansion.

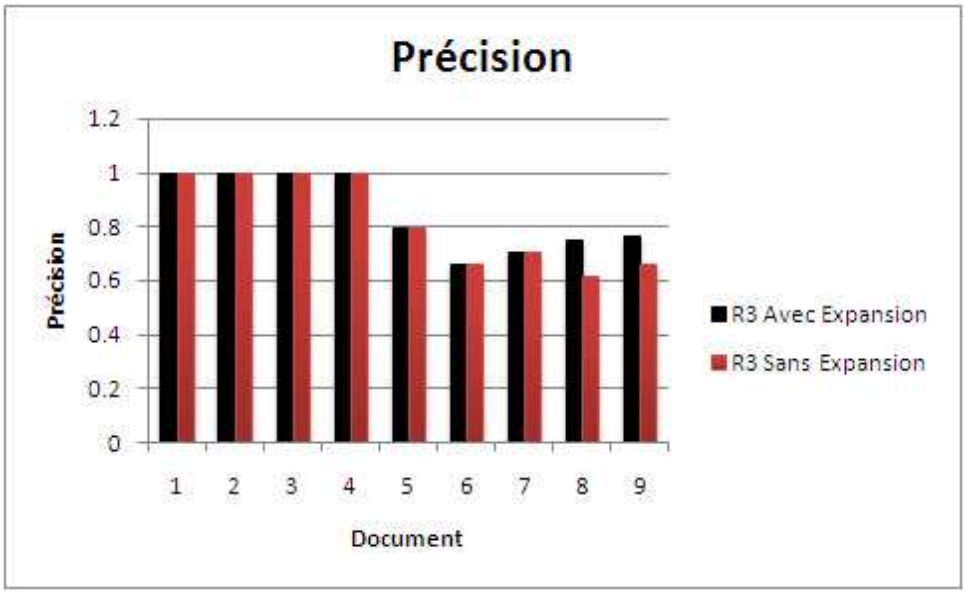

Fig. 9.2. Niveau Précision pour la requête R_3 avec et sans expansion.

La Figure 9.3 représente la comparaison des courbes Rappel/Précision pour la requête R3 avec et sans expansion.

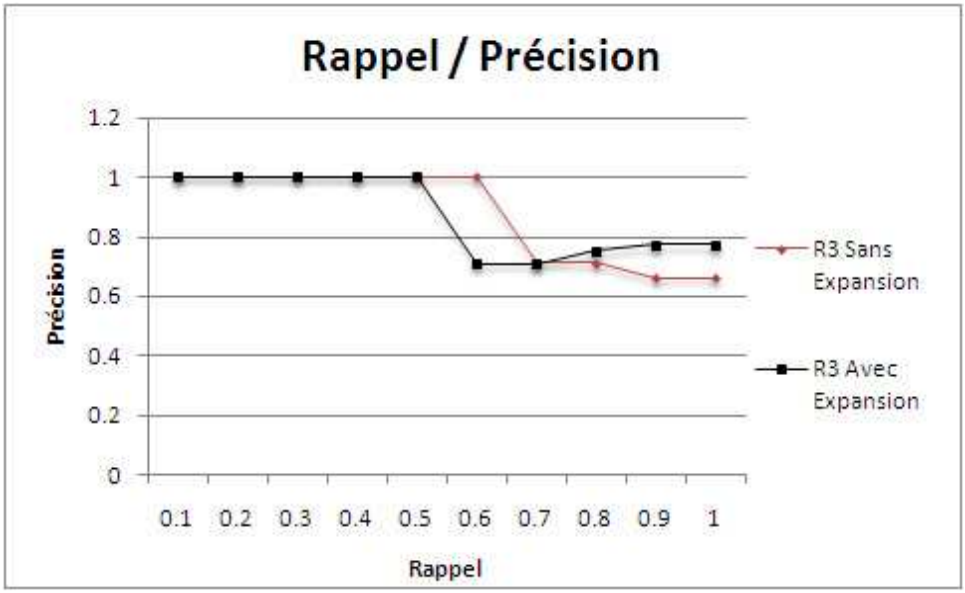

Fig. 9.3. Courbes Rappel/Précision pour la requêtes R_3 avec et sans expansion.

Les niveaux de précision pour les rappels de 0,8, 0.9 et 1 sont supérieurs de 4%, 11% et 11% respectivement pour la requête R_3 avec expansion par rapport à la requête R_3 sans expansion. Des hausses comparables sont constatées pour les autres critères d'évaluation : amélioration de 3 % de précision moyenne et dégradation de 6% de rappel moyen pour la requête R_3 avec expansion par rapport à la requête R_3 sans expansion. Les faibles écarts entre les résultats mentionnés montrent que la technique de retour de pertinence testée a légèrement contribué pour l'amélioration des performances du système de recherche d'information.

3. La segmentation thématique pour le ré-ordonnancement des résultats

Généralement, un système de recherche d'information (SRI) retourne, en réponse à la requête d'un utilisateur, une liste de documents ordonnée selon leurs degrés de pertinence. Pour permettre à l'utilisateur d'un système de recherche d'information d'accéder plus rapidement à ce qu'il cherche, on peut se limiter de ne lui proposer que les passages pertinents d'un document et non pas le document dans son intégralité. Il est donc possible de voir la recherche d'information comme une tâche devant rapporter des zones textuelles pertinentes (Passage Retrieval) et non pas des documents entiers. Pour extraire les unités documentaires pertinentes, une segmentation thématique des documents est évidemment nécessaire. Cette segmentation permet la mise en valeur de certains documents grâce à un nouveau calcul des similarités avec la requête à partir du découpage des textes en unités documentaires. Elle permet également de réordonner la liste originelle de réponses grâce à la correspondance entre ces unités documentaires et les documents auxquels elles appartiennent. Pour cela, les segments thématiques sont indexés comme si elles étaient des documents à part entière. La seconde recherche fonctionne selon le même procédé que la première.

3.1. Evaluation de la segmentation thématique au sein d'un SRI

Dans ce qui suit, la segmentation sera évaluée en utilisant deux types corpus (C_1 et C_2) avec un jeu de six (06) requêtes différentes pour chaque corpus. Le Corpus C_1 est une collection de 280 documents scientifiques avec un total de 1107 unités documentaires. Le corpus C_2 est ensemble de 340 traditions prophétiques contenant 1065 unités documentaires (Voir Chapitre V, Section 5.2 pour plus de détails). Les deux jeux de requête sont données dans les tableaux 9.4 et 9.5 ci-dessous.

Numéro de la requête	Texte de la requête
1	التدخين، أسبابه، نتائجه و مصادره
2	كيف كانت الحياة الأدبية في العصر العباسي
3	أمراض العصر، الأمراض القاتلة و كذا المعدية
4	تأثير الكحول على المعدة
5	المشتري
6	اختراعات المصباح و الاتصال

Tableau 9.4. Jeu de requêtes utilisé pour l'évaluation du Corpus C_1.

Numéro de la requête	Texte de la requête
1	عذاب الميت ببكاء أهله عليه
2	الطواف بين الصفا و المروة
3	كيف فرضت الصلاة في الإسراء
4	ما تعلمه سيدنا موسى من الخضر
5	صلاة الرسول صلى الله عليه و سلم
6	كيف كان سؤال هرقل عن النبي صلى الله عليه و سلم

Tableau 9.5. Jeu de requêtes utilisé pour l'évaluation du Corpus C_2.

3.1.1. Exemple de segmentation

Pour la requête n°6 du premier jeu de requêtes, le document rapporté est découpé en 3 segments. Les termes de la requête sont: «اختراعات, المصباح, الاتصال ». Le document trouvé en réponse à la requête évoque l'histoire de l'ampoule. Il est segmenté en trois segments correspondant aux thématiques traitées dans le document.

Segment n° 1 : (مصباح الفتيلة)

... وفي عام 1784 اخترع الكيميائي السويسري (ايميه ارغاند) مصباح ذا فتيلة أنبوبية و ركب عليها مدخنة من أجل توجيه الهواء نحو الشعلة و بالتالي زيادة فعالية المصباح الزيتي المستخدم في العصر الحجري ، وفي عام 1799 سجلت أول براءة اختراع في باريس لمصباح يعمل على حرق الغاز ...

Segment n° 2 : (المصباح الكهربائي)

... لنعود الآن إلى العام 1878م انكب أديسون على العمل في مختبره يحدوه الأمل في التوصل إلى المصباح الكهربائي ،آنذاك شاع خبر أن أديسون يريد إضاءة العالم، وبدأت الصحف بنشر الخبر وكان تعليقها هو أن الأمر هذا فوق طاقة البشر ، لكن أديسون استمر بالعمل مع (40) عاملاً ليل نهار في (مينلوبارك) و يبذلون جهدهم في سبيل تحقيق الهدف الذي يعتقد العلم استحالة الوصول إليه...

Segment n° 3 : (ثورة الكهرباء)

... كان المصباح الكهربائي حجر الأساس إلى ثورة الكهرباء التي نعيش في أحضانها حتى اليوم. وقد سعى أديسون إلى بناء محطات لتوليد الكهربا وتوزيعها على البيوت لاستخدامها في إنارة مصباحه، والطريف في الأمر أن المحطة الأولى بنيت في بريطانيا وعرفت باسم محطة هولبورن فيادلت المركزية وقد تم افتتحها رسمياً في 13-يناير 1882م لإمداد عدة شوارع ومباني قريبة بالكهرباء...

Fig. 9.4. Segmentation thématique du document D_1 en réponse à la Requête R_6 du premier jeu de requêtes.

Le deuxième exemple de segmentation thématique concerne un document qui a été rapporté par le système de recherche d'information en réponse à la requête n°4 du deuxième jeu de requêtes. Ce document est découpé en 2 segments. Les termes de la requête sont: « ما, تعلمه, سيدنا, موسى, من, الخضر ».

Segment n° 1 : (عتاب الله لموسى عليه السلام)

... قام موسى النبي خطيبا في بني إسرائيل فسئل: أي الناس أعلم؟ فقال: أنا أعلم، فعتب الله عليه، إذ لم يرد العلم إليه، فأوحى الله إليه: إن عبدا من عبادي بمجمع البحرين، هو أعلم منك. قال: يا رب، وكيف به؟ فقيل له: احمل حوتا في مكتل، فإذا فقدته فهو ثم، فانطلق وانطلق بفتاه يوشع بن نون، وحمل حوتا في مكتل، حتى كانا عند الصخرة وضعا رؤوسهما وناما، فانسل الحوت من المكتل فاتخذ سبيله في البحر سربا، وكان لموسى وفتاه عجبا، فانطلقا، بقية ليلتهما ويومهما، فلما أصبح قال موسى لفتاه: آتنا غداءنا لقد لقينا من سفرنا هذا نصبا. ولم يجد موسى مسا من النصب حتى جاوز المكان الذي أمر به، قال له فتاه: أرأيت إذ أوينا إلى الصخرة؟ فإني نسيت الحوت، قال موسى: ذلك ما كنا نبغي، فارتدا على آثارهما قصصا، فلما انتهيا إلى الصخرة، إذا رجل مسجى بثوب، أو قال تسجى بثوبه، فسلم موسى، فقال الخضر: وأنى بأرضك السلام؟ فقال: أنا موسى، فقال: موسى بني إسرائيل؟ قال: نعم، قال: هل أتبعك على أن تعلمني مما علمت رشدا؟ قال: إنك لن تستطيع معي صبرا، يا موسى، إني على علم من علم الله علمنيه لا تعلمه أنت، وأنت على علم علمكه لا أعلمه. قال: ستجدني إن شاء الله صابرا، ولا أعصي لك أمرا.

Segment n° 2 : (تعلّم موسى من الخضر)

فانطلقا يمشيان على ساحل البحر، ليس لهما سفينة، فمرت بهما سفينة، فكلموهم أن يحملوهما، فعرف الخضر، فحملوهما بغير نول، فجاء عصفور فوقع على حرف السفينة، فنقر نقرة أو نقرتين في البحر، فقال الخضر: يا موسى: ما نقص علمي وعلمك من علم الله إلا كنقرة هذا العصفور في البحر، فعمد الخضر إلى لوح من ألواح السفينة فنزعه، فقال موسى: قوم حملونا بغير نول، عمدت إلى سفينتهم فخرقتها لتغرق أهلها؟ قال: ألم أقل لك إنك لن تستطيع معي صبرا؟ قال: لا تؤاخذني بما نسيت - فكانت الأولى من موسى نسيانا - فانطلقا، فإذا غلام يلعب مع الغلمان، فأخذ الخضر برأسه من أعلاه فاقتلع رأسه بيده، فقال موسى: أقتلت نفسا زكية بغير نفس؟ قال: ألم أقل لك إنك لن تستطيع معي صبرا؟ - قال ابن عيينة: وهذا أوكد - فانطلقا، حتى إذا أتيا أهل قرية استطعما أهلها فأبوا أن يضيفوهما، فوجد فيها جدارا يريد أن ينقض فأقامه، قال الخضر بيده فأقامه، فقال له موسى: لو شئت لاتخذت عليه أجرا، قال: هذا فراق بيني وبينك). قال النبي صلى الله عليه وسلم: (يرحم الله موسى، لوددنا لو صبر حتى يقص علينا من أمرهما)...

Fig. 9.5. Segmentation thématique d'un document D_2 en réponse à la Requête R_4 du deuxième jeu de requêtes.

Les tableaux 9.6 et 9.7 montrent les résultats d'évaluation du système de recherche d'information pour les deux corpus C_1 et C_2 sans utilisation de la technique de segmentation thématique. Les nombres de documents trouvés pour chaque requête à partir de chaque corpus sont aussi indiqués.

Requête	NDT	NDPT	NDP	Rappel	Précision
1	1	1	2	0.50	1.00
2	12	3	3	1.00	0.25
3	15	4	4	1.00	0.26
4	5	1	1	1.00	0.20
5	2	2	2	1.00	1.00
6	13	3	4	0.75	0.23

Tableau 9.6. Résultats d'évaluation du Corpus C_1 (sans segmentation).

Requête	NDT	NDPT	NDP	Rappel	Précision
1	2	1	1	1.00	0.50
2	2	2	5	0.40	1.00
3	3	1	1	1.00	0.33
4	2	1	1	1.00	0.50
5	36	6	6	0.16	1.00
6	36	1	1	1.00	0.02

Tableau 9.7. Résultats d'évaluation du Corpus C_2 (sans segmentation).

Les tableaux 9.8 et 9.9 montrent les résultats d'évaluation du système de recherche d'information pour les deux corpus C_1 et C_2 avec utilisation de la technique de segmentation thématique. Ces tableaux indiquent le nombre de segments thématiques trouvés pour chaque requête à partir des deux corpus C_1 et C_2. Pour les requêtes (R_2 et R_3) du premier jeu de requêtes et les requêtes (R_4 et R_5) du deuxième jeu de requêtes, aucune segmentation n'a été obtenue.

Requête	Titre	NS_D	NST	NSP	NSP_D	Rappel	Précision
1	التدخين في الجزائر	7	1	1	3	0.33	1.00
4	القرحة المعدية	3	1	1	1	1.00	1.00
5	المشتري	1	1	1	1	1.00	1.00
6	المصباح	10	3	3	5	0.60	1.00

Tableau 9.8. Résultats d'évaluation du Corpus C_1 (avec segmentation).

Requête	Titre	NS_D	NST	NSP	NSP_D	Rappel	Précision
1	البكاء على الميت	4	2	2	2	1.00	1.00
2	حجة النبي ﷺ	11	1	1	1	1.00	1.00
3	فرض الصلاة في الإسراء	5	1	1	4	0.25	1.00
4	سؤال العالم أي الناس أعلم	4	2	2	2	1.00	1.00

Tableau 9.9. Résultats d'évaluation du Corpus C_2 (avec segmentation).

Avec: NDT: le nombre de documents trouvés. NDPT: le nombre de documents pertinents trouvés. NDP: le nombre de documents pertinents. Titre: le titre du document évalué. NSD : le nombre de segments par document. NST: le nombre de segments trouvés. NSP : le nombre de segments pertinents. NSPD : le nombre de segments pertinents par document.

3.1.2. Résultats d'évaluation

Ces résultats sont obtenus en mesurant les similarités entre les segments thématiques et la requête selon le même model exposé dans le chapitre V (model vectoriel et mesure Cosine). Le score final du document correspond au score des segments les plus proches de la requête.

Les figures 9.6, 9.7, 9.8, 9.9 et 9.10 montrent la comparaison des niveaux de rappel, de précision et de la mesure F_1 du système de recherche d'information avec et sans utilisation de la technique de segmentation thématique pour les deux corpus C_1 et C_2 respectivement.

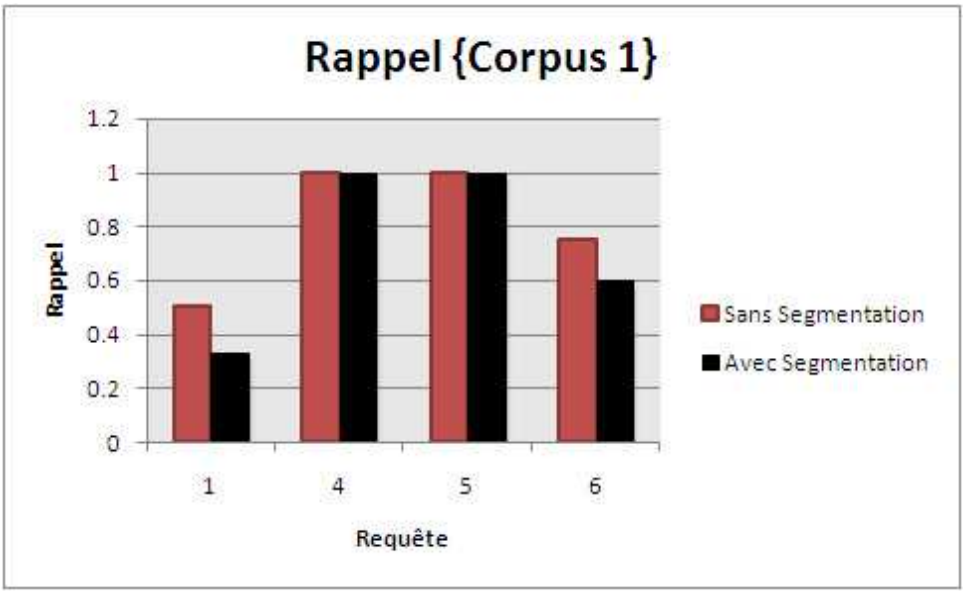

Fig. 9.6. Niveau de Rappel pour le corpus C_1 avec et sans segmentation.

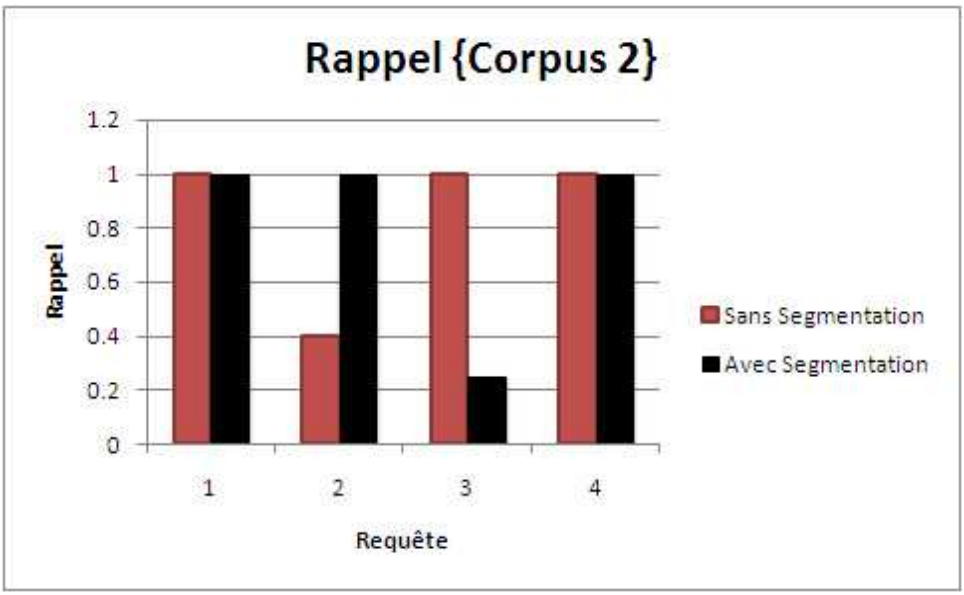

Fig. 9.7. Niveau de Rappel pour le corpus C_2 avec et sans segmentation.

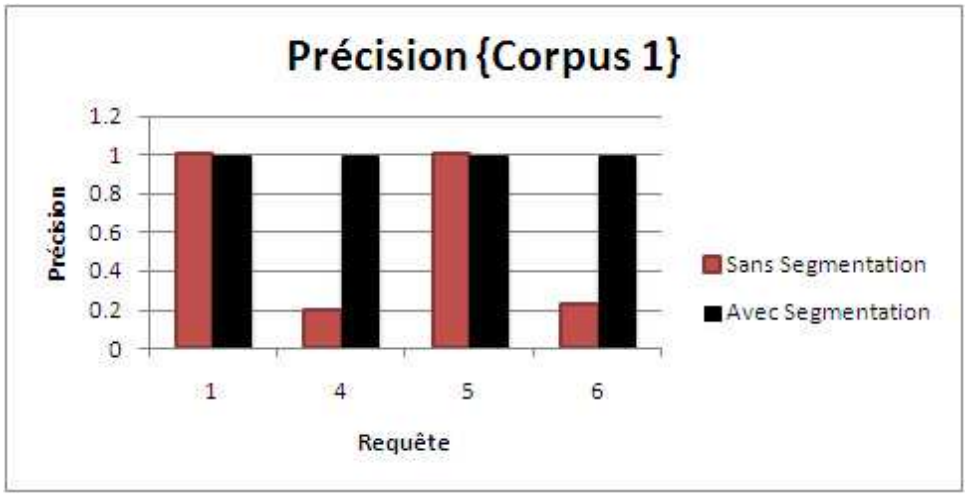

Fig. 9.8: Niveau de Précision pour le corpus C_1 avec et sans segmentation.

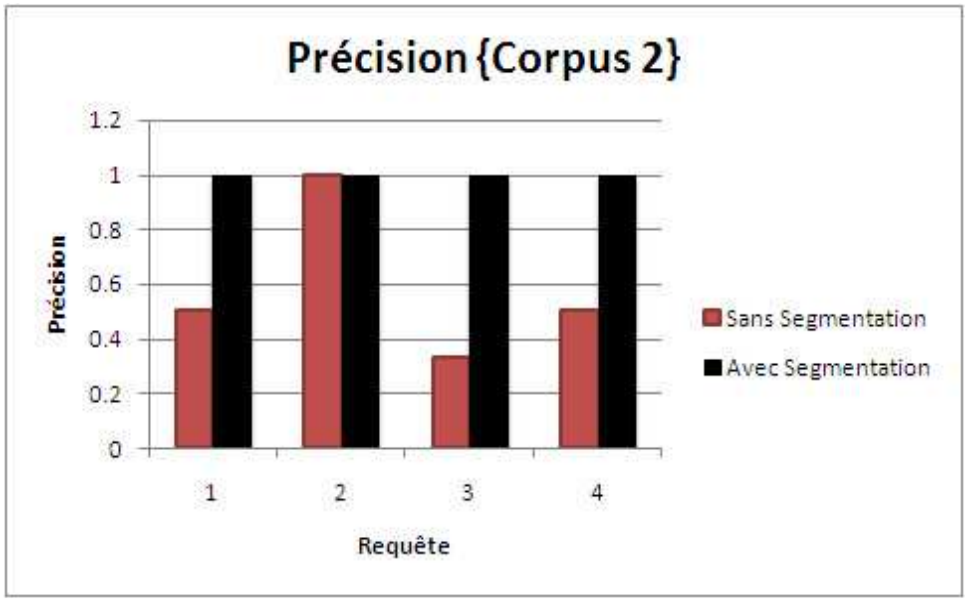

Fig. 9.9. Niveau de Précision pour le corpus C_2 avec et sans segmentation.

Lorsque le score final d'un document correspond au score du meilleur segment, une baisse absolue de 17%, de 15% et de 75% du rappel par rapport à une utilisation sans segmentation a été constatée pour les requêtes 1 et 6 pour le corpus C_1 et la requête 3 pour le corpus C_2. On a aussi constaté une amélioration de 80 % et de 77 % absolus pour les précisions respectifs des requêtes 4 et 6 du premier jeu de requêtes et de 50 %, de 67 % et de 50 % absolus pour les précisions respectifs des requêtes 1, 3 et 4 du deuxième jeu de requêtes par rapport à une utilisation sans segmentation.

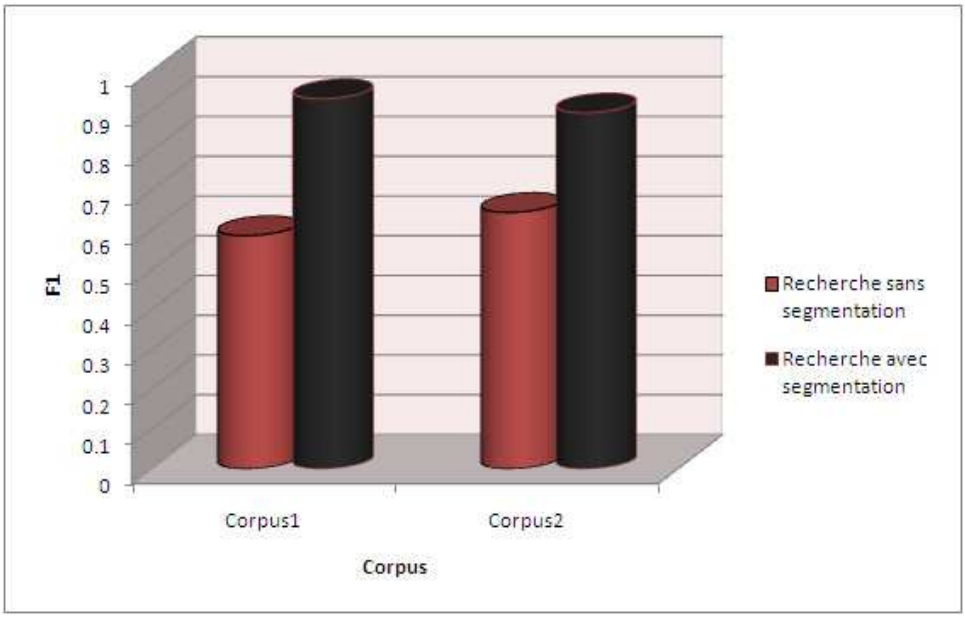

Fig. 9.10. Comparaison des deux corpus.

La figure 9.11 et 9.12 montre la comparaison des niveaux de la mesure F_1 et des courbes Rappel/Précision du système de recherche d'information avec et sans utilisation de la technique de segmentation thématique pour le requête n=°4 du premier jeu de requêtes.

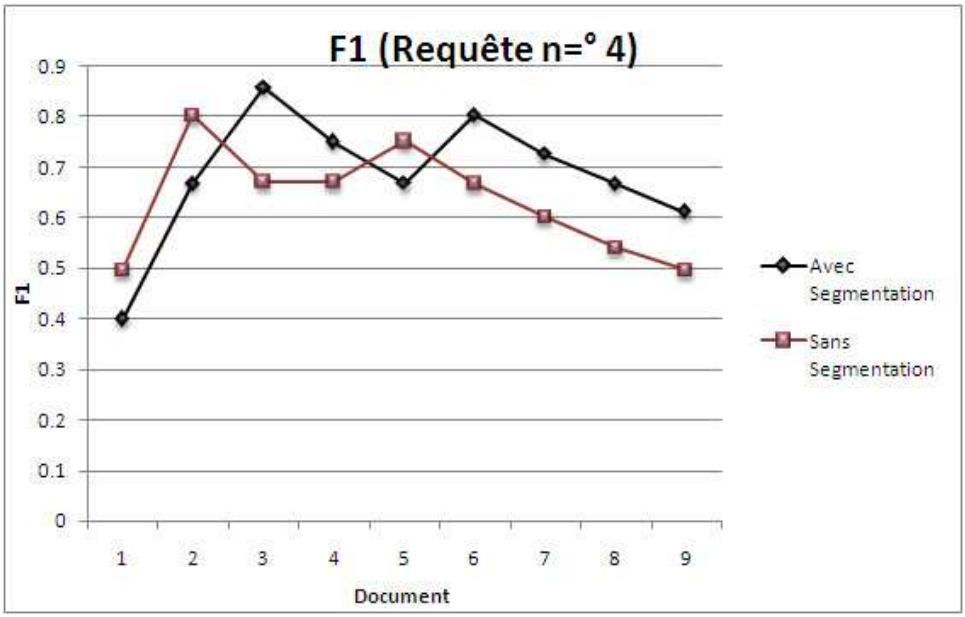

Fig. 9.11. Comparaison des valeurs F1 pour la Requête n=°4 du corpus C_1.

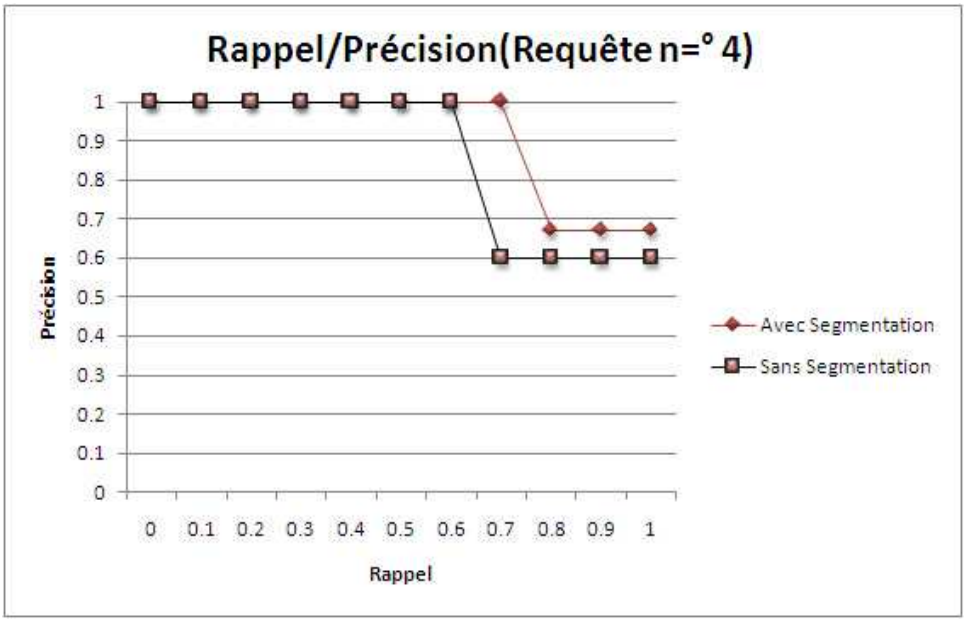

Fig. 9.12. Comparaison des courbe Rappel/précision pour la Requête n=°4 du corpus C_1.

D'après les résultats d'évaluation présentés dans la figure 9.10, l'amélioration de la mesure F1 par rapport à l'utilisation sans segmentation est de 34 % et de 25 % absolus pour les corpus C_1 et C_2 respectivement. Le fait de considérer les passages des documents permet donc, tel que mentionné dans de nombreux travaux, une réorganisation linéaire des documents réponse de manière efficace. Les meilleures améliorations sont à noter pour le corpus C_1, où les passages sont très hétérogènes.

D'après les résultats d'évaluation présentés dans les figures 9.11 et 9.12, l'amélioration de la mesure F1 par rapport à l'utilisation sans segmentation pour la requête n=° 4 est entre 3 % et 9 % absolus pour les 5 derniers documents. Par contre, une baisse de 17 %, de 13 % et de 6 % de la mesure F1 est constatée pour les trois premiers documents D_1, D_2 et D_3 respectivement. Si l'on analyse les résultats des courbes rappel/précision pour la même requête, la segmentation permet d'améliorer le niveau de précision de 40 % pour un rappel de 0.7. Le niveau de précision des rappels 0.8, 0.9 et 1 est amélioré d'une valeur sensible de 7 %.

3.2. Résultats et discussion

Pour étudier l'influence de la tâche de segmentation thématique sur les performances de notre système de recherche d'information, l'ensemble des segments trouvés est considéré comme un nouveau corpus (les segments sont traités comme s'ils étaient des documents entiers). La fonction de recherche des segments est effectuée selon le même model que la fonction de recherche des documents (les valeurs de pondération des termes sont différentes en raison de la nouvelle segmentation). À la fin de cette deuxième recherche, deux listes peuvent être proposés à l'utilisateur. La première est la liste des segments thématique classés en fonction de leur similarité avec la requête tandis que la seconde est la liste des documents originaux classés en fonction des résultats de leurs segments (la nouvelle valeur de pertinence d'un document correspond au score de son meilleur segment). Une étude particulière doit être faite pour trouver la bonne fonction de similarité permettant de mesurer la ressemblance d'une requête avec un segment thématique d'un document qui est parfois très court. Comme suggéré dans [Bellot et El-Bèze 2001], une combinaison linéaire des différentes mesures donne de bons résultats. La combinaison des scores obtenus par les segments d'un document avec son score initial permet de prendre en compte la pertinence globale d'un document et d'être plus en adéquation avec la nature du corpus.

Les résultats de notre évaluation, montrent que l'utilisation des fragments de texte (segments thématiques) pour estimer la proximité des documents à la requête permet d'améliorer l'efficacité des systèmes de recherche d'information. Les documents peuvent en effet contenir des passages traitant de sujets très divers. Le fait de considérer ces passages séparément permet de donner plus de poids à des zones de texte pertinentes noyées dans une masse d'informations sans rapport avec le besoin exprimé par l'utilisateur. Dans les approches dites de *Passage Retrieval*, où les similarités sont calculées indépendamment pour chaque passage d'un ensemble de documents, un document qui ne possède qu'une petite partie de texte correspondant au sujet de l'utilisateur a alors plus de chances d'être examiné.

4. La classification thématique pour l'organisation de l'information pertinente

Offrant des alternatives plus qu'intéressantes à la classique liste ordonnée de documents, les techniques de classification thématique permettent bien souvent de réduire les efforts à fournir pour localiser l'information recherchée. L'idée est alors de réaliser une recherche d'information préliminaire en considérant la requête de l'utilisateur et d'appliquer le processus de classification sur le sous-ensemble ainsi obtenu [Preece, 1973]. Ici, l'objectif est d'améliorer l'interprétation des résultats, de réduire les efforts cognitifs que l'utilisateur doit fournir pour localiser les informations qu'il recherche. Non seulement la catégorisation des résultats d'une recherche d'information classique permet d'orienter l'utilisateur vers les documents pertinents plus rapidement mais cela peut aussi le renseigner sur la diversité des informations du corpus en rapport avec son sujet [Lampier 2008].

En effet, les bénéfices résultant de la catégorisation des documents pour leur présentation à l'utilisateur proviennent du fait que l'identification des documents pertinents n'est plus basée sur les seuls termes de la requête mais tire profit des relations existant entre les documents considérés. Par exemple, un document contenant un faible nombre de termes en commun avec la requête peut quand même être identifié comme pertinent grâce sa proximité à d'autres documents pertinents. [Lampier 2008].

L'objectif final est d'extraire les parties les plus intéressantes d'un ensemble de documents retournés par un système de recherche d'information afin de présenter à l'utilisateur une liste de passages de texte lui permettant de sélectionner les groupes de passages thématiques, qui lui semblent correspondre au mieux à ses besoins. La figure 9.13 ci-dessous présente le modèle général de notre approche.

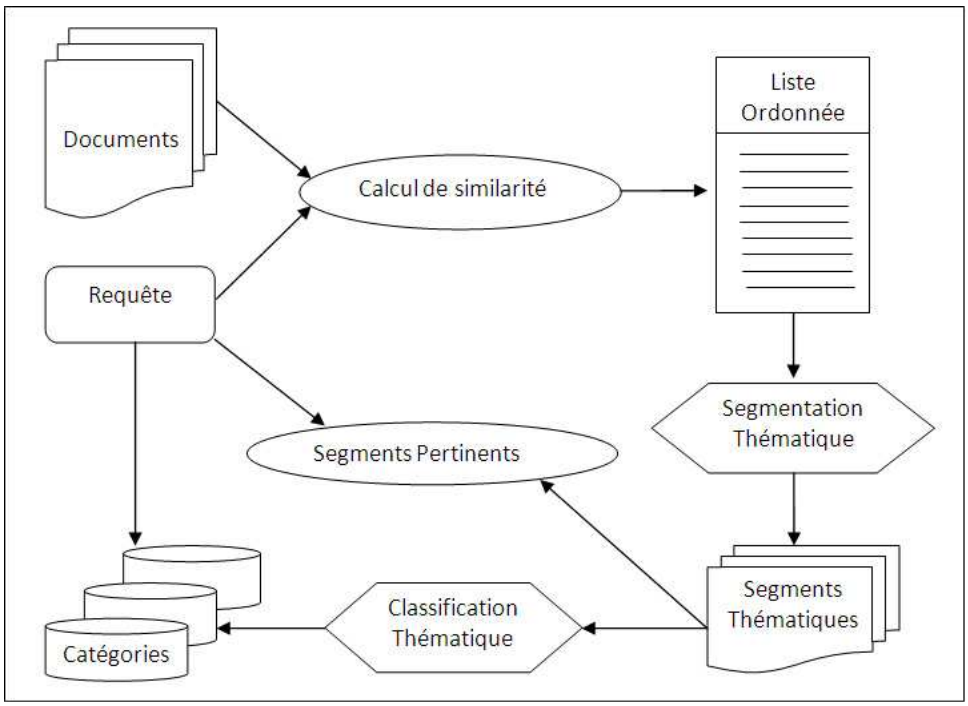

Fig. 9.13. Modèle proposé pour l'organisation de l'information pertinente.

4.1. Evaluation de la classification thématique au sein d'un SRI

Pour l'évaluation de l'impact de la technique de classification thématique sur les performances de notre système de recherche d'information, nous utilisons le même jeu de requêtes utilisé dans la section 2.2. Les hadiths pertinents à la requête R_3 (H_{5834}) après la première recherche sont présentés dans le tableau 9.10 ci-dessous.

Kitab	Bab	Num Hadith	Matn Hadith
المناقب	كنية النبي صلى الله عليه وسلم	3345	**تسموا باسمي ولا تكتنوا بكنيتي**
المناقب	كنية النبي صلى الله عليه وسلم	3346	**سموا باسمي ولا تكتنوا بكنيتي**
الأدب	من سمى بأسماء الأنبياء. وقال أنس: قبل النبي صلى الله عليه وسلم إبراهيم، يعني ابنه. [ر:1241]	5843	**سموا باسمي ولا تكتنوا بكنيتي**، فإنما أنا **قاسم أقسم** بينكم
المناقب	كنية النبي صلى الله عليه وسلم	3344	كان النبي صلى الله عليه وسلم في السوق، فقال رجل: يا **أبا القاسم**، فالتفت النبي صلى الله عليه وسلم، فقال: (**سموا باسمي، ولا تكنوا بكنيتي**)
الأدب	قول النبي صلى الله عليه وسلم: (سموا باسمي ولا تكنوا بكنيتي). قاله أنس، عن النبي صلى الله عليه وسلم	5833	ولد لرجل منا غلام **فسماه القاسم**، فقالوا: لا **نكنيه** حتى نسأل النبي صلى الله عليه وسلم، فقال: (**سموا باسمي ولا تكتنوا بكنيتي**)
البيوع	ما ذكر في الأسواق	2014	كان النبي صلى الله عليه وسلم في السوق، فقال رجل: يا **أبا القاسم**، فالتفت إليه النبي صلى الله عليه وسلم، فقال: إنما دعوت هذا، فقال النبي صلى الله عليه وسلم: (**سموا باسمي، ولا تكنوا بكنيتي**)
البيوع	ما ذكر في الأسواق	2015	دعا رجل بالبقيع: يا **أبا القاسم**، فالتفت إليه النبي صلى الله عليه وسلم فقال: لم

أعنك، قال: (**سموا باسمي ولا تكنوا بكنيتي**)			
ولد له غلام، فأراد أن **يسميه** محمدا، قال: (**سموا باسمي، ولا تكنوا بكنيتي، فإني إنما جعلت قاسما أقسم بينكم**). وقال حصين: (**بعثت قاسما أقسم بينكم**).	2946	باب: قول الله تعالى: {فأن لله خمسه وللرسول} /الأنفال: 41/. يعني: للرسول قسم ذلك، قال رسول الله صلى الله عليه وسلم: (إنما أنا قاسم وخازن، والله يعطي)	الخمس
ولد لرجل منا غلام **فسماه القاسم**، فقالت الأنصار: لا **نكنيك أبا القاسم** ولا ننعمك عينا، فأتى النبي صلى الله عليه وسلم فقال: يا رسول الله، ولد لي غلام، **فسميته القاسم**، فقالت الأنصار: لا **نكنيك أبا القاسم** ولا ننعمك عينا، فقال النبي صلى الله عليه وسلم: (**أحسنت الأنصار، سموا باسمي ولا تكنوا بكنيتي، فإنما أنا قاسم**).	2947	باب: قول الله تعالى: {فأن لله خمسه وللرسول} /الأنفال: 41/. يعني: للرسول قسم ذلك، قال رسول الله صلى الله عليه وسلم: (إنما أنا قاسم وخازن، والله يعطي)	الخمس
تسموا باسمي ولا تكتنوا بكنيتي، ومن رآني في المنام فقد رآني حقا، فإن الشيطان لا يتمثل في صورتي، ومن كذب علي متعمدا فليتبوأ مقعده من النار	110	إثم من كذب على النبي صلى الله عليه وسلم	العلم
سموا باسمي ولا تكنوا بكنيتي، ومن رآني في المنام فقد رآني، فإن الشيطان لا يتمثل صورتي، ومن كذب علي متعمداً فليتبوأ مقعده من النار	5844	من سمى بأسماء الأنبياء	الأدب
ولد لرجل منا غلام **فسماه القاسم**، فقلنا: لا **نكنيك أبا القاسم** ولا كرامة، فأخبر النبي صلى الله عليه وسلم فقال: (**سم** ابنك عبد الرحمن)	5832	أحب الأسماء إلى الله عز وجل	الأدب
ولد لرجل منا غلام **فسماه القاسم**، فقالوا: لا **نكنيك بأبي القاسم** ولا ننعمك عيناً، فأتى النبي صلى الله عليه وسلم فذكر ذلك له، فقال: (**أسم** ابنك عبد الرحمن)	5835	أحب الأسماء إلى الله عز وجل	الأدب
أن أباه جاء إلى النبي صلى الله عليه وسلم فقال: (ما **اسمك**). قال: حزن، قال: (أنت سهل). قال: لا أغير **اسماً سمانيه** أبي.	5836	اسم الحزن	الأدب
... قال: (ما **اسمه**). قال: فلان، قال: (ولكن **اسمه** المنذر). **فسماه** يومئذ المنذر	5838	تحويل الاسم إلى اسم أحسن منه	الأدب
أخنى **الأسماء** يوم القيامة عند الله رجل **تسمى** ملك الأملاك	5852	أبغض الأسماء إلى الله	الأدب

Tableau 9.10. Liste globale des réponses pour la requête R_3 (H_{5834}).

La liste des segments pertinents à la requête R_3 (H_{5834}) après application de la segmentation thématique à l'ensemble des réponses est donnée dans le tableau 9.11 ci-dessous. Les segments sont classés selon leur appartenance aux catégories thématiques du corpus. Les résultats sont obtenus en mesurant les similarités entre les segments thématiques et la requête selon le même model exposé dans le chapitre 3 (model vectoriel et mesure Cosine). Le score final du document correspond au score des segments les plus proches de la requête.

Catégorie	Hadith	Segment Pertinent	Similarité
المناقب	3345	تسموا باسمي ولا تكتنوا بكنيتي	1.00
المناقب	3346	سموا باسمي ولا تكتنوا بكنيتي	1.00
المناقب	3344	سموا باسمي، ولا تكنوا بكنيتي	1.00
البيوع	2014	سموا باسمي، ولا تكنوا بكنيتي	1.00
البيوع	2015	سموا باسمي ولا تكنوا بكنيتي	1.00
الخمس	2946	سموا باسمي، ولا تكنوا بكنيتي، فإني إنما جعلت قاسما أقسم بينكم	0.88
الخمس	2947	أحسنت الأنصار، سموا باسمي ولا تكنوا بكنيتي، فإنما أنا قاسم	0.93
العلم	110	تسموا باسمي ولا تكتنوا بكنيتي، ومن رآني في المنام فقد رآني حقا، فإن الشيطان لا يتمثل في صورتي، ومن كذب علي متعمدا فليتبوأ مقعده من النار	0.70
الأدب	5833	سموا باسمي ولا تكتنوا بكنيتي	1.00
الأدب	5843	سموا باسمي ولا تكتنوا بكنيتي، فإنما أنا قاسم أقسم بينكم	0.89
الأدب	5844	سموا باسمي ولا تكنوا بكنيتي، ومن رآني في المنام فقد رآني، فإن الشيطان لا يتمثل صورتي، ومن كذب علي متعمداً فليتبوأ مقعده من النار	0.70
الأدب	5832	سم ابنك عبد الرحمن	0.41
الأدب	5835	أسم ابنك عبد الرحمن	0.41

Tableau 9.11. Liste des réponses pour la requête R_3 (H_{5834}) après expansion.

Pour effectuer une telle évaluation, nous devrions connaitre à priori la véritable catégorie thématique de chacun des groupes de segments. Pour ce faire, nous avons décidé d'attribuer comme véritable étiquette thématique de chaque groupe le nom de la catégorie thématique des segments constituant le plus grand pourcentage du groupe. La figure 9.14 montre la répartition des segments pertinents par catégorie thématique.

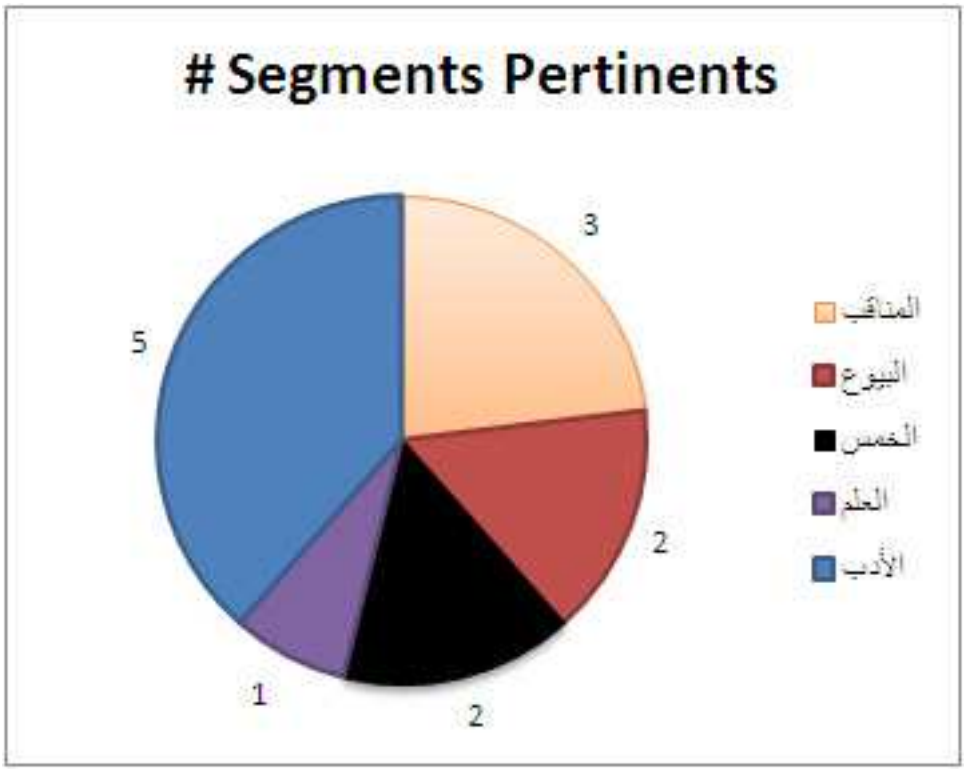

Fig. 9.14. répartition des segments pertinents par catégorie thématique..

La figure 9.15 montre l'évaluation des scores de similarité de chaque catégorie thématique avec la requête R_3 (H_{5834}). Le score de chaque catégorie thématique est obtenu en calculant la moyenne de scores des segments pertinents appartenant à cette catégorie.

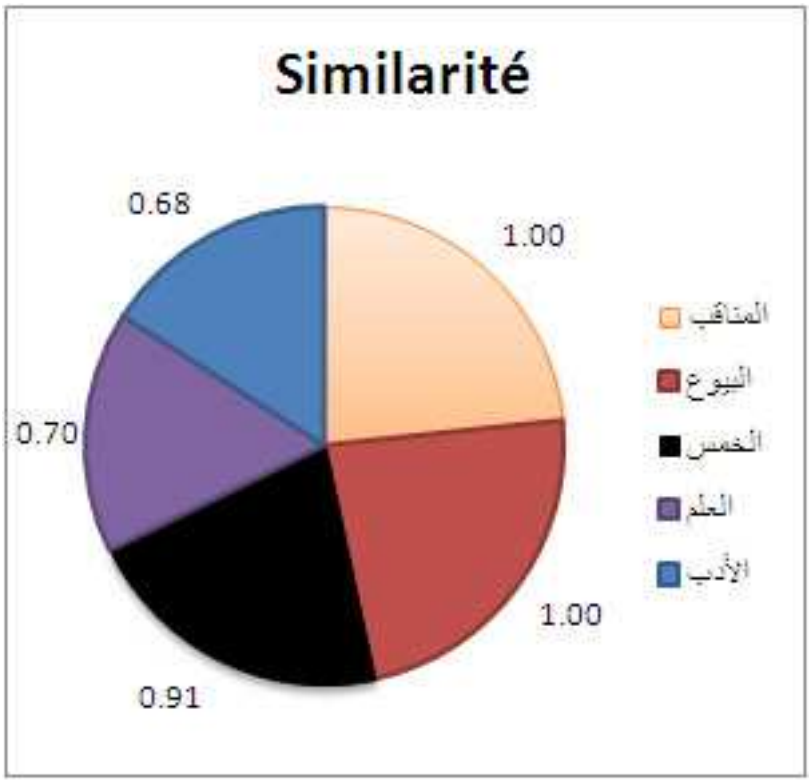

Fig. 9.15. l'évaluation des scores de similarité de chaque catégorie thématique avec la requête R_3 (H_{5834}).

Le tableau 9.12 indique les résultats d'évaluation obtenus sur notre corpus prophétique selon la R_3 requête après classification thématique.

Catégorie	Doc	Pertinence	Précision	Rappel
المناقب	3345	1	1.00	0.11
المناقب	3346	1	1.00	0.22
المناقب	3344	1	1.00	0.33
العلم	110	1	1.00	0.44
الأدب	5833	1	1.00	0.55
الأدب	5843	1	1.00	0.66
الأدب	5844	1	1.00	0.77
الأدب	5832	1	1.00	0.88
الأدب	5835	1	1.00	1.00

Tableau 9.12. Résultats d'évaluation pour la requête R_3 après classification thématique.

Les Figures 9.16 et 9.17 représentent la comparaison des niveaux de Rappel et de Précision pour la requête R_3 avec et sans classification thématique. La Figure 9.18 représente la comparaison des courbes Rappel/Précision pour la requête R3 avec et sans classification thématique.

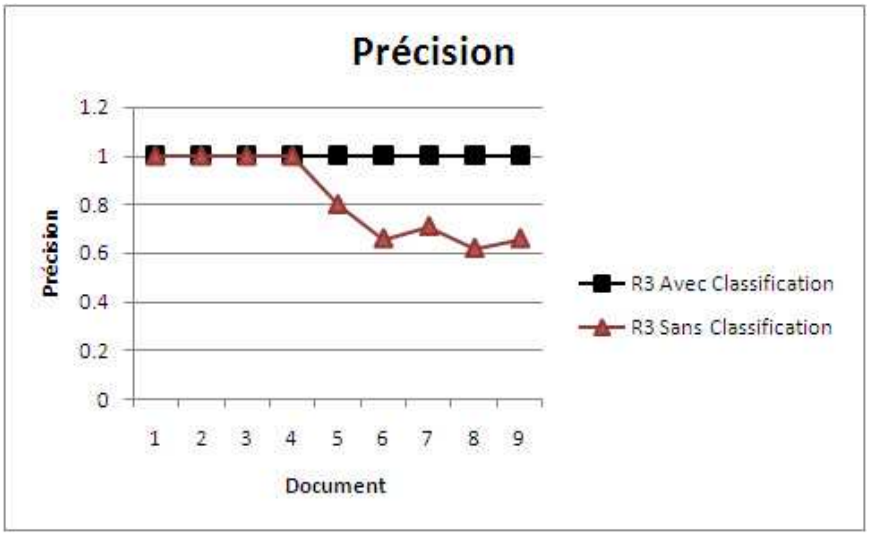

Fig. 9.16. Niveau de Rappel pour la requête R_3 avec et sans classification.

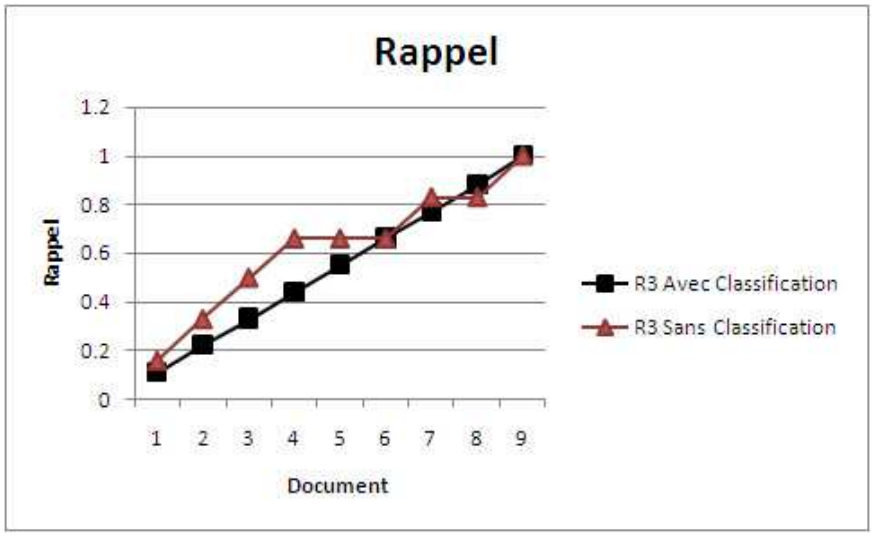

Fig. 9.17. Niveau de Rappel pour la requête R_3 avec et sans classification.

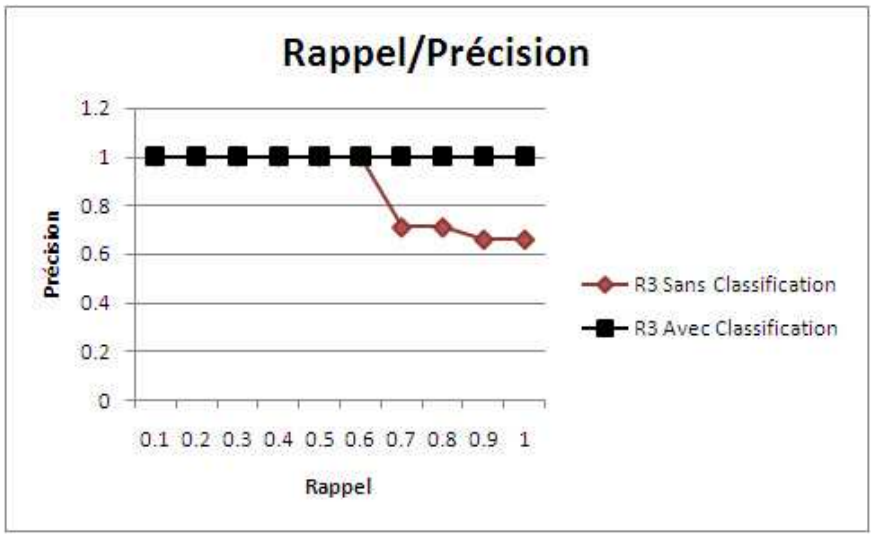

Fig. 9.18. Courbes Rappel/Précision pour la requêtes R_3 avec et sans classification.

Lorsque le score final d'un document correspond au score du meilleur groupe thématique des segments pertinents, une baisse absolue du rappel par rapport à une utilisation sans classification thématique a été constatée pour les rappels des documents de D_1 à D_5 et le rappel du document D_7. On a aussi constaté une amélioration de 20 %, 34 %, 29 %, 38%, et de 34 % absolus pour les précisions respectifs des documents D_5 à D_9 par rapport à une utilisation sans classification. Si l'on analyse les résultats des courbes rappel/précision pour la même requête, la classification permet d'améliorer le niveau de précision de 29 %, de 29%, de 34 % et de 34 % pour les niveaux de rappels 0.7, 0.8, 0.9 et 1 respectivement.

4.2. Résultats discussion

Dans cette section, nous avons expérimenté l'impact de l'utilisation des segments thématiques des documents sur un processus de catégorisation thématique appliqué aux résultats d'une recherche d'information. L'objectif est de considérer les différents thèmes de chaque document séparément pour produire des groupes de documents plus cohérents. Même si les similarités n'ont pas été nécessairement améliorées par la considération de tels passages, les catégories obtenues tendent à être plus représentatives des thèmes abordés par les documents. Les expériences réalisées en recherche d'information ont montré que le découpage thématique des documents permet d'obtenir des catégories thématiques contenant une meilleure proportion de documents pertinents. Les expériences réalisées ont montré qu'un découpage thématique des documents permettait d'obtenir des catégories présentant un meilleur pourcentage de documents pertinents, et ainsi de fournir à l'utilisateur un outil d'aide pour la tâche d'accès aux informations pertinentes.

5. Conclusion

Ce chapitre a été consacré à l'étude et l'évaluation des techniques utilisées pour l'organisation de l'information pertinentes afin d'améliorer l'accès à l'ensemble des résultats retournés par système de recherche d'information.

D'après les résultats de la première partie de ce chapitre, on a conclus que la reformulation de requêtes est une phase très importante pour les systèmes de recherche d'information. Elle permet en effet de réécrire la requête de l'utilisateur selon les informations retrouvées par la requête initiale. De manière générale, ceci consiste, dans le cas du retour de pertinence, d'extraire à partir des documents jugés pertinents par l'utilisateur, les mots clés importants puis les rajouter à la requête.

Les expérimentations menées dans la deuxième partie de ce chapitre, ont visé à comparer l'utilisation de la recherche d'information classique par rapport à la recherche d'information utilisant la segmentation thématique comme une base pour la fourniture de nouveaux fonctionnalités d'organisation et d'accès à l'information pertinente.

L'évaluation effectuée concerne la capacité des méthodes de segmentation à mettre en valeur certains documents à partir de leur segmentation. Les exemples des documents segmentés illustrent bien la capacité de la méthode à isoler certaines parties thématiquement homogènes des documents. Cette évaluation donne des résultats acceptables.

Dans ce chapitre nous nous sommes intéressés à la réorganisation des réponses fournies par un SRI. Nous avons montré que le fait de considérer des passages thématiques peut s'avérer utile pour collecter de l'information pertinente. Nous avons aussi montré que le regroupement des segments thématiques selon leur catégories thématiques, permet de fournir à l'utilisateur un aperçu des différentes thématiques qu'il peut trouver dans le corpus, en rapport avec une requête.