Utilisation de Séquences Fréquentes Maximales en Recherche d'Information

Antoine Doucet[1,2]
doucet@cs.helsinki.fi, doucet@info.unicaen.fr

[1]Department of Computer Science - PO Box 26 - FIN-00014 University of Helsinki - Finlande
[2]Université de Caen - Département d'Informatique - Campus Côte de Nacre - F-14032 Caen Cedex - France

Abstract

The growing amount of textual information electronically available has increased the need for high precision retrieval. The use of phrases was long seen as a natural way to improve retrieval performance over the common document models that ignore the sequential aspect of word occurrences in documents, considering them as *bags of words*. However, both statistical and syntactical phrases showed disappointing results for large document collections.

In this paper we present a new kind of phrases in the form of *Maximal Frequent Sequences*. Rather mined phrases than statistical phrases, their main strengths are to form a very compact index and to account for the sequentiality and adjacency of meaningful word co-occurrences. They also allow a gap between words.

We introduce a method for using these phrases in information retrieval and present our experiments on the INEX collection, a 494 Mb collection of scientific articles. When aggregating the retrieved documents using our phrases with the results of our baseline vector space model system, its average precision for the best 100 documents was improved by 22.8%. The state of the art gives much weaker improvements for similar-sized document collections.

Résumé

La quantité croissante de données textuelles sous forme électronique a augmenté l'importance de la précision des systèmes de recherche d'information. L'utilisation de phrases a toujours été vue comme une technique naturelle pour améliorer la performance des systèmes de recherche. Les techniques classiques sont en effet basées sur des modèles de documents ne tenant pas compte de l'ordre des mots qui les composent. Un document est alors considéré comme un *sac de mots*. Cependant, les phrases statistiques et syntaxiques ont jusqu'ici obtenu des résultats décevants pour des collections de grande taille.

Dans cet article, nous présentons les *Séquences Fréquentes Maximales* (SFM), une nouvelle sorte de phrases basées sur des fréquences d'occurrence, mais plutôt issues de techniques de fouille de texte que d'analyse statistique pure. Ces phrases indexent une collection de documents de façon très compacte. En outre, elles tiennent compte de la séquencialité des mots et de leur adjacence, autorisant même un gap entre deux mots formant une même séquence.

Nous introduisons une méthode pour exploiter ces phrases en recherche d'information et présentons nos expériences sur la collection INEX, composée de 12,107 articles scientifiques pour une taille totale de 494 Mo. En aggrégeant les résultats obtenus à l'aide des séquences fréquentes maximales à ceux obtenus en utilisant une technique standard, nous améliorons la précision moyenne des 100 premières réponses de 22,8%. L'état de l'art présente des résultats beaucoup plus faible pour des collections de documents d'une taille similaire.

Keywords: Statistical Phrases, Syntactical Phrases, Information Retrieval, Text Mining.

1 Introduction

Le nombre croissant de documents électroniques rend nécessaires des systèmes de recherche d'information de plus en plus précis. La précision d'un système étant le pourcentage de documents pertinents parmi le nombre total de documents répondus à une requête.

La majorité des systèmes de recherche d'information ne tiennent pas compte de l'ordre des mots dans un document. Afin d'améliorer leurs performances, il est raisonnable de penser qu'il existe des solutions basées sur cette prise en compte. Zhai et al. [12] évoquent différents types de problèmes causés par l'utilisation de mots simples. Ils constatent notamment que certaines associations de mots ont un sens différent de «l'addition» des sens de ces deux mots (e.g., l'expression «cordon bleu» ne désigne généralement pas un cordon qui est bleu). Les expressions métaphoriques posent un problème similaire (e.g., «avoir un chat dans la gorge»).

Des travaux sur l'utilisation de phrases en recherche d'information ont lieu depuis plus de 25 ans. Les premiers résultats fûrent très prometteurs. Toutefois, de façon inattendue, la constante augmentation de la taille des collections de documents utilisées a induit une baisse drastique de la qualité des résultats. En 1975, Salton et al. [7] rapportent en effet une amélioration de la précision moyenne sur 10 points de rappel comprise entre 17% et 39%. En 1989, Fagan [2] réitère les mêmes expériences avec une collection de 10 Mo et obtient des améliorations de 11% à 20%. Cet impact négatif de la taille de la collection fût dernièrement confirmé par Mitra et al. [5], qui pour une collection de 655 Mo n'améliore plus la précision que d'un pourcent ! Turpin et Moffat [9] revisitent encore ces expériences en 1999 et obtiennent des améliorations de précision incluses entre 4% et 6%.

Une conclusion de ces travaux est que les phrases améliorent les résultats aux bas niveaux de rappel, mais sont globalement inefficaces pour les n premiers documents retournés. D'après Mitra et al. [5], cet apport moindre des phrases pour les meilleures réponses s'expliquent par le fait que leur utilisation promotionne des documents n'évoquant qu'un seul aspect d'une requête. Par exemple, partant d'une requête portant sur les problèmes associés aux fonds de pension, beaucoup des réponses les mieux classés évoquent les fonds de pension, mais aucune difficulté associée. Le problème se rapporte à celui d'une *couverture inadéquate* de la requête.

A notre sens, ceci ne remet pas en cause l'idée qu'ajouter des descripteurs tenant compte de l'ordre des mots doit permettre d'améliorer la performance des systèmes de recherches d'information. Mais ces travaux mettent en évidence le besoin de combiner différemment l'apport des phrases et celui des mots simples [8] et surtout le fait que les phrases extraites selon les techniques actuelles ne sont pas satisfaisantes dans l'objectif de représenter les documents d'une collection.

Dans la deuxième section, nous décrivons brièvement le modèle d'espace vectoriel (aussi nommé «sac de mots»), puis présentons les différents types de phrases utilisées dans des travaux liés. En section 3, nous définissons la notion de séquence fréquente maximale et présentons l'algorithme d'extraction correspondant. Nous présentons ensuite la technique de traitement des requêtes correspondante visant à tirer profit des SFM dans le

cadre applicatif de la recherche d'information (section 4), avant de présenter notre cadre expérimental et les résultats obtenus (section 5). Nous terminons cet article en tirant les conclusions et en dressant les perspectives prochaines de ce travail (section 6).

2 Utilisation de phrases en recherche d'information

2.1 Description du modèle d'espace vectoriel

Représentation des Documents. La représentation par le modèle d'espace vectoriel est la plus usitée. Chaque document d'une collection y est représenté par un vecteur à N dimensions, où N correspond au nombre de *caractéristiques* décrivant la collection. Dans la plupart des approches, les caractéristiques sont les mots les plus significatifs.

Un vecteur représentant un document contient le poids de chaque caractéristique dans ce document. Une valeur fréquemment utilisée pour ce poids est *tf-idf*. *Tf-idf* est une combinaison du nombre d'occurrences du terme dans le document (tf signifie «term frequency») et de la valeur inverse du nombre de documents dans lesquels il est présent (idf signifie «inverted document frequency»).

La mesure du nombre d'occurrences d'un terme dans la collection (tf) ne permet pas de capturer sa spécificité. Or un terme commun à de nombreux documents est moins utile qu'un terme commun à peu d'entre eux. C'est ce qui motive la combinaison des mesures tf et idf. En bref, tf mesure l'importance d'une terme dans un document, tandis qu'idf mesure sa spécificité dans une collection.

Mesure de Similarité Etant donné une requête et une collection de documents, afin d'obtenir un ensemble de réponses, il est nécessaire de comparer la requête aux documents. Dans le cadre du modèle d'espace vectoriel, on représente la requête par un vecteur appartenant au même espace. Le modèle d'espace vectoriel prend alors tout son sens : il est possible d'utiliser des techniques d'algèbre simple pour calculer des mesures de similarité entre les documents. La mesure la plus fréquente est le cosinus, dont l'atout principal est sa faible complexité. En effet, lorsque les vecteurs sont normalisés, $cosinus(\overrightarrow{d_1}, \overrightarrow{d_2})$ se simplifie en $(d_1 \cdot d_2)$.

2.2 Utilisation de phrases

Il existe pour cela différentes possibilités. La méthode usuelle est de considérer les phrases comme des dimensions supplémentaires de l'espace vectoriel, au même titre que les mots simples. Cela pose toutefois quelques problèmes. Le poids donné aux phrases les moins fréquentes est faible. Leur spécificité est pourtant souvent décisive pour déterminer la pertinence d'un document. L'interdépendance entre les différents termes est également problématique. Comment tenir compte du lien entre le poids d'une séquence et celui des deux mots qui la composent ?

Il existe principalement 2 types de phrases, les phrases statistiques, issues d'un simple comptage des co-occurrences de mots seuls, et les phrases syntaxiques.

Phrases Statistiques. Pour Mitra et al. [5], une phrase statistique est formée pour chaque paire de 2 mots lemmatisés adjacents et apparaissant dans au moins 25 documents de la collection TREC-1. Les paires sont ensuite triées par ordre lexicographique. Nous relevons ici au moins 2 problèmes. Premièrement, ce classement lexicographique revient à ignorer une information séquentielle précédemment découverte sur une paire de mots : son ordre ! Cela revient à dire que AB=BA. En outre, aucun gap n'est autorisé, il est cependant fréquent de représenter un même concept en ajoutant un mot entre deux autres. Cette définition de phrase ne constate par exemple aucune similarité entre les deux fragments de texte «Université de Basse-Normandie» et «Université de Caen Basse-Normandie». Ce modèle est très éloigné de la réalité du langage naturel.

Phrases Syntaxiques. La méthode d'extraction de phrase syntaxique de Mitra et al. utilise la nature et la fonction des mots et accepte comme phrases syntaxiques toutes les séquences maximales de mots acceptées par une grammaire découlant d'un ensemble de motifs prédéfinis. Par exemple, une séquence «verbe, nombre cardinal, adjectif, adjectif, nom commun» constituera une phrase de taille 5. Toutes les sous-paires occurrant dans cet ordre seront également générées, avec un gap illimité (e.g., la paire «verbe, non commun» sera générée). Cette technique permet de très bien représenter le langage naturel. Malheureusement, obtenir nature et fonction des mots est très coûteux. La taille de l'index est également conséquente, puisque toutes les phrases sont stockées, quel que soit leur nombre d'occurrence. Dans les expériences, Mitra reconnaît en fait ne pas créer d'index pour la collection à priori, mais générer les phrases en réaction à une requête. En pratique, cela signifie un temps d'attente très important pour l'utilisateur. Les résultats sont pourtant similaires à ceux obtenus avec les phrases statistiques.

Nous supposons que cela est dû à plusieurs facteurs. Le premier est certainement l'absence d'un seuil de fréquence minimal pour indexer une phrase. Cela signifie que des phrases très rares ont une influence majeure sur les résultats, alors que leur rareté peut simplement témoigner d'une anomalie. Autoriser un gap illimité pour générer les sous-paires paraît également dangereux : la phrase «I like to eat hot dogs» générera la paire «hot dogs», mais aussi la paire «like dogs», dont le sens sémantique n'a rien à voir avec celui de la phrase initiale.

Séquences Fréquentes Maximales. Nous proposons donc les SFM comme une alternative afin de tenir compte de l'ordre des mots dans la modélisation de documents textuels. Elles présentent l'avantage de n'être extraites que si elles apparaissent avec une fréquence minimale (supérieure à un seuil donné), évitant ainsi l'extraction de phrases non significatives. Un gap entre deux mots est également autorisé au sein même du processus d'extraction, permettant d'appréhender une plus grande variété d'expression.

3 Séquences Fréquentes Maximales

La technique d'extraction des *Séquences Fréquentes Maximales* (SFM) [1] d'une collection de documents inclut trois étapes principales. L'idée générale respecte les principes de

la fouille de données ; sélection et élagage, puis application des techniques centrales du processus de fouille, suivi d'une dernière phase dont le but est de transformer les résultats en connaissances compréhensibles.

3.1 Définition

Définition 1. Une séquence $p = a_1...a_k$ est une *sous-séquence* d'une séquence q si tous les items a_i, $1 \leq i \leq k$, occurrent dans q et qu'ils occurrent dans le même ordre que dans p. Si une séquence p est une sous-séquence d'une séquence q, on dit alors que p *occurre* dans q.

Définition 2. Une séquence p est *fréquente* dans une collection de documents D si p est une sous-séquence occurrant dans au moins σ documents de D, où σ est un seuil de fréquence documentaire donné.

Définition 3. Une séquence p est une *(sous-)séquence fréquente maximale* de D s'il n'existe pas de séquence $p' \in D$ telle que p soit une sous-séquence de p', et que p' soit fréquente dans D.

D'après les définitions précédentes, une séquence est dite maximale si et seulement si aucune autre séquence fréquente ne contient cette séquence.

3.2 Algorithme d'Extraction

Prétraitement. Cette étape préalable consiste à «nettoyer» les données. Les caractères spéciaux (incluant, par exemple, la ponctuation et les parenthèses) sont effacés. Pour éviter de traiter des items inintéressants, un antidictionnaire est utilisé. Il contient articles, pronoms, conjonctions, adverbes communs, et les formes fréquentes des verbes non-informatifs (e.g., «est», «a», «es»). Ces éléments sont ignorés.

Phase initiale : Collection des paires fréquentes. Cette phase sert à collecter toutes les paires de mots dont la fréquence documentaire est supérieure à un seuil donné σ. Deux mots forment une paire s'ils occurrent dans le même document, et si la distance qui les séparent est inférieure à un *gap* g donné. Notons également que les paires sont ordonnées, ce qui signifie que les paires (A,B) et (B,A) sont distinctes.

Expansion des paires fréquentes. Pour chaque étape k, $Grams_k$ est le nombre d'ensembles fréquents de longueur k. Ainsi, les paires fréquentes calculées durant la phase initiale composent $Grams_2$. Les SFMs sont trouvées en combinant les séquences fréquentes courtes (de taille k) dans le but de former des séquences plus longues (de taille $k+1$). Chaque étape inclut de nombreuses phases d'élagage, pour tenter de contrer les risques d'explosion combinatoire.

Au terme du processus, chaque document de la collection est décrit par un ensemble de SFMs.

3.3 Principaux Atouts de la Méthode

La technique permet d'extraire toutes les séquences fréquentes maximales de mots d'une collection de documents. En outre, un *gap* entre les mots est autorisé. Dans une phrase, les mots n'ont pas besoin d'apparaître de façon continue : un paramètre g indique combien d'autres mots deux mots d'une séquence peuvent avoir entre eux. Ce paramètre g est normalement choisi entre 1 et 3.

Par exemple, si g = 2, une phrase «président Bush» est trouvée dans chacun des 2 fragments textuels suivants :
`...Le président des Etats Unis George Bush...`
`...Président George W. Bush...`
Note : Les articles et les prépositions ont été supprimés durant le prétraitement.

L'autorisation d'un gap entre les mots d'une séquence est probablement le plus grand atout de cette méthode, comparée aux autres méthodes existantes pour l'extraction de descripteurs textuels. Cela augmente grandement la qualité des phrases, puisque ce traitement prend en compte la variété du langage naturel.

L'autre spécificité avantageuse des SFMs est la possibilité d'extraire des séquences fréquentes maximales de n'importe quelle taille. Cela permet une description de documents très compacte. Par exemple, en plafonnant la longueur des phrases à 8, une séquence fréquente de mots de longueur 25 nécessiterait plusieurs milliers de phrases pour être représentée (ce qu'un simple calcul combinatoire permet de vérifier aisément).

3.4 Une technique efficace pour extraire une approximation

Malheureusement, la présence du gap implique un coup computationnel non négligeable. L'algorithme est exponentiel en le nombre de documents et en leur taille. En pratique, cette exponentialité signifie que pour certains corpus, les SFMs sont calculées en quelques secondes, tandis que pour d'autres collections plus grandes, elles ne peuvent pas être extraites en pratique.

Pour résoudre ce problème, nous avons développé une technique qui permet d'extraire une approximation de l'ensemble des SFMs d'une collection de documents. En divisant la collection de documents en une partition de plusieurs sous-collections, en extrayant l'ensemble des SFM pour chaque collection, et finalement, en joignant chacun de ces ensembles de SFM, nous obtenons une approximation de l'ensemble des SFM de la collection complète.

La validité de cette technique repose sur la conjecture que les SFM extraites sont issues de documents similaires par nature, et qu'en groupant les documents similaires entre eux, la perte d'information induite par ce partitionnement intermédiaire devrait être minimale. Pour former les sous-collections de documents, nous avons utilisé l'algorithme de clustering *k-means* [11], qui présente l'avantage d'être de complexité linéaire.

En utilisant de petites collections de documents, nous avons pu vérifier la qualité de cette approximation de façon empirique. En ce qui concerne la computabilité, le résultat est

net : pour la collection INEX utilisée dans nos expériences, il s'est tout simplement avéré impossible d'extraire les SFM par la méthode directe. En utilisant notre technique de partitionnement intermédiaire, nous avons obtenu des résultats en quelques heures sur un ordinateur familial.

4 Traitement des requêtes

4.1 Discussion et Objectifs

Etant donné un ensemble de séquences décrivant les documents d'une collection, comment déterminer dans quelle mesure une séquence $p_1 \dots p_n$ décrivant une collection de documents D correspond à une séquence $q_1 \dots q_m$ trouvée dans une requête correspondante ? Et comment conséquemment établir un classement des documents supposés les plus pertinents relativement à cette requête ?

Notre approche consiste à extraire un ensemble de séquences fréquentes décrivant chaque document de la collection. Ces séquences fréquentes sont ensuite comparées aux phrases-clés trouvées dans la requête de l'utilisateur. Chaque document reçoit une *quantité de pertinence* pour chaque séquence qu'il contient correspondant à une phrase de la requête. Ce bonus peut être différent pour chaque phrase.

Il est en effet notamment souhaitable de favoriser les phrases dont l'usage est plus spécifique, en utilisant des coefficients statistiques tenant compte de leur fréquence et de leur spécificité.

Il est également naturel de supposer qu'une plus grande quantité de pertinence découle d'une correspondance plus longue. Si une requête contient la phrase «recherche d'information structurée en XML», il est naturel de privilégier les phrases contenant cette séquence exacte, puis celles en contenant une sous-séquence de taille 3 (par exemple, «recherche d'information structurée»), et enfin celles contenant une sous-séquence de taille 2 (par exemple, «recherche d'information» ou «information structurée»).

Il apparaît également utile de tenir compte du fait que le langage naturel est moins rigide que ne l'est la définition des séquences fréquentes maximales. Il ne serait en effet pas raisonnable d'ignorer la similitude entre une requête ABC et une séquence CBA ou même CAB. On souhaitera cependant prendre en compte le fait que la similarité entre ABC et CAB est plus forte que celle entre ABC et CBA.

Dans l'esprit des séquences fréquentes maximales, nous souhaitons aussi concrétiser numériquement la notion de gap. La phrase AC contient ainsi généralement une forte similitude sémantique avec la phrase ABC (par exemple «information XML structurée» et «information structurée»), quoique dans le cas général, cette similitude avec la phrase ABC est moins forte que celle qui lie la phrase ABC aux paires AB et BC.

Dans la prochaine sous-section, nous présentons la technique que nous proposons et qui tient compte de toutes ces observations.

4.2 Méthode

Dans une première étape, nous extrayons les séquences fréquentes descriptives des documents d'une collection suivant la technique décrite en section 3. Pour pouvoir comparer les séquences de mots représentant les documents et les phrases-clés issues d'une requête, nous décomposons les séquences en paires. Ceci permet de comparer des objets de même taille et d'obtenir ainsi des mesures de similarité cohérentes. Chaque paire issue d'une phrase-clé est associée à un score représentant sa *quantité de pertinence*. Cette quantité de pertinence représente «l'importance» de la présence d'une paire de mots dans les documents correspondants. Cette valeur est modifiée par un coefficient d'adjacence qui réduit la quantité de pertinence attribuée par une paire formée de deux mots qui n'apparaissent pas côte à côte dans la phrase-clé dont ils sont issus.

4.2.1 Définitions :

Soient D une collection de N documents et $A_1 \dots A_n$ une phrase de n mots issus d'une requête contre la collection D. La quantité de pertinence associée à la paire de mots A_iA_j est donnée par :

$$Q_{pertinence}(A_iA_j) = idf(A_iA_j, D) \cdot adj(A_iA_j)$$

où $idf(A_iA_j, D)$ représente la spécificité de A_iA_j dans la collection D :

$$idf(A_iA_j, D) = \log\left(\frac{N}{n}\right)$$

et $adj(A_iA_j)$ est le coefficient d'adjacence visant à pénaliser les paires de mots composées de mots non adjacents dans $A_1 \dots A_n$:

$$adj(A_iA_j) = \begin{cases} 1, & \text{si } A_i \text{ et } A_j \text{ sont adjacents} \\ 0 \leq \alpha_1 \leq 1, & \text{si d}(A_i,A_j) = 1 \\ 0 \leq \alpha_2 \leq \alpha_1 & \text{si d}(A_i,A_j) = 2 \\ \dots & \\ 0 \leq \alpha_{n-2} \leq \alpha_{n-3}, & \text{si d}(A_i,A_j) = n-2 \end{cases}$$

D'évidence on souhaite que $(i \geq j) \Rightarrow (\alpha_j \geq \alpha_i)$, c'est à dire qu'une plus grande distance entre 2 mots implique une moindre pertinence pour la paire correspondante. Dans les expériences, nous nous limiterons à une distance de 1 (i.e., $\forall k > 1 : \alpha_k = 0$).

Notons que la valeur d'adjacence de A_iA_j dans $A_1 \dots A_n$ est aussi nommée le *coefficient modificateur de* A_iA_j.

4.2.2 Exemple :

En ignorant les distances supérieures à 1, une phrase-clé ABCD d'une requête sera décomposée en 5 couplets (paire, coefficient modificateur) :

Document	SFM	Paires correspondantes	Matchs	Quantité de pertinence
d_1	AB	AB	AB	idf(AB)
d_2	ACD	AC CD AD	AC CD	idf(CD) + α_1.idf(AC)
d_3	AFB	AF FB AB	AB	idf(AB)
d_4	ABC	AB BC AC	AB BC AC	idf(AB) + idf(BC) + α_1.idf(AC)
d_5	ACB	AC CB AB	AC AB	idf(AB) + α_1.idf(AC)

Tab. 1 – Quantité de pertinence de différentes phrases contre une requête ABCD

(AB, 1), (BC, 1), (CD, 1), (AC, α_1), (BD, α_1)

Comparons cette requête aux documents d_1, d_2, d_3, d_4 et d_5, respectivement décrits par les séquences fréquentes AB, AC, AFB, ABC et ACB (notons ici que le fait que chaque document soit décrit par une seule phrase est un cas particulier qui ne vaut qu'à titre d'exemple). Les quantités de pertinence apportées par la requête ABCD sont indiquées dans la table 1.

On constate que les coefficients modificateurs forment bien un ordre du type souhaité et décrit précédemment. Le seul manque notable est la non prise en compte des paires apparaissant dans l'ordre inverse de celui de la requête (e.g. pour la requête ABCD : BA).

4.3 Aggrégation des scores de similarité

Il est évident qu'en pratique, de nombreuses requêtes ne contiennent pas de phrase, et que certaines phrases donneront peu ou pas de résultats. En outre, les documents contenant les mêmes phrases obtiennent tous le même score. Il faut donc les départager afin de pouvoir décider d'un ordre de présentation des résultats à l'utilisateur.

Une idée naturelle serait de re-décomposer les paires en mots simples et de comparer ces mots à ceux de la requête. Cela n'est cependant pas satisfaisant, car les mots les moins fréquents ne peuvent être extraits par l'algorithme d'extraction des séquences fréquentes maximales. Une catégorie de mots potentiellement non extraits, et plus importante encore, est celle des mots qui sont fréquents mais qui ne co-occurrent pas fréquemment avec d'autres.

Pour remplir ce manque, nous extrayons séparemment une valeur de pertinence pour les mots seuls, suivant le modèle d'espace vectoriel présenté dans la sous-section 2.1. Les caractéristiques utilisées sont les mots simples, à l'exception de ceux supposés les moins informatifs, écartés lors d'une phase de prétraitement : petits adverbes, verbes auxiliaires, articles, etc. Il reste alors à combiner les scores de pertinence obtenu par les SFM et ceux obtenus par les mots simples. Pour cela, nous devons d'abord les rendre comparables en les ramenant sur le même intervalle [0,1] grâce à Max_Norm présenté par Lee [4] :

$$Nouveau\ Score = \frac{Ancien\ Score}{Score\ Maximal}$$

A l'issue de cette première étape, nous combinons les nouveaux scores normalisés en

utilisant un facteur d'interpolation linéaire λ, représentant le poids relatif des réponses données par chacune des 2 techniques [10].

$$Score\ Aggrege = \lambda \cdot score_{Mots\ Simples} + (1 - \lambda) \cdot score_{SFM}$$

Nos expériences nous ont donné de bons résultats en donnant le nombre de mots simples distincts **issus de phrases-clés** de la requête comme poids du score des mots simples, et le nombre de mots simples distincts des phrases de la requête comme poids du score issus des SFM.

5 Expériences et Résultats

Nous avons basé nos expériences sur la collection de documents d'INEX (Initiative for the Evaluation of XML retrieval). INEX a vu le jour en 2002 pour répondre à la demande des chercheurs en recherche d'information structurée qui ne bénéficiaient pas jusqu'alors d'un forum d'évaluation spécifique. Une spécificité des documents de la collection INEX est qu'ils ont une structure logique XML fournie. Dans les expériences présentées, nous n'utilisons cependant pas cette structure. La collection INEX se compose d'environ 12,107 articles scientifiques en anglais de l'IEEE, ainsi que d'un ensemble de requêtes et d'assessements compulsés manuellement par les participants. Ces assessements manuels nous permettent d'évaluer numériquement les résultats de notre système.

5.1 Indexation

Le premier traitement est d'enlever tous les éléments de ponctuation, les chiffres, les mots de moins de trois lettres, ainsi que ceux appartenant à l'antidictionnaire. Nous extrayons alors les phrases en utilisant un seuil de fréquence de 7, soit la plus petite valeur permettant de calculer l'ensemble des séquences fréquentes maximales en un temps raisonnable.

5.2 Traitement des requêtes

Nous n'avons utilisé que les 25 requêtes d'INEX 2002 ne faisant pas cas de la structure XML des documents (CO) et comportant des assessements manuels complets. De ces requêtes, nous n'avons utilisé que les mots et phrases-clés situés dans la balise «Keywords». Un exemple de contenu d'une telle balise est celui de la requête 47 :

```
 <Keywords>
"concurrency control" "semantic transaction management" "application"
"performance benefit" "prototype" "simulation" "analysis"
</Keywords>
```

Dans le modèle d'espace vectoriel, nous avons calculé la similarité entre cette requête et les documents de la collection (i.e., le cosinus des vecteurs correspondants) et classé les documents par score de similarité décroissant.

précision@n	Mots simples	SFM	Aggrégé	Amélioration Aggrégé/Mots Simples
précision@10	0,62886	0,53693	0,59933	**-4,7%**
précision@50	0,60918	0,42478	0,58163	**-4,5%**
précision@100	0,05296	0,03467	0,06506	**+22,8%**

Tab. 2 – Précision stricte moyenne pour les n premiers documents retournés

Pour traiter une requête en utilisant les SFM, nous avons décomposé chaque phrase de la requête en paires comme décrit dans la section 4, en utilisant la valeur d'adjacence arbitraire α_1=0,8 (faire varier ce paramètre n'a qu'une très faible incidence sur les résultats).

Pour obtenir les scores aggrégés, nous calculons λ en fonction du nombre de mots simples total (11 dans l'exemple) et du nombre de mots simples occurrant dans une phrase-clé (7 dans l'exemple). Pour la requête présentée ci-dessus, cela donne $\lambda = \frac{11}{11+7}$.

5.3 Résultats

La précision moyenne en considérant les 10, 50 et 100 premiers résultats retournés est donnée en table 2. Cette valeur est obtenue de façon classique à partir d'une courbe de rappel-précision obtenue elle-même suivant la méthode décrite par Raghavan et al. [6] et reprise par Gövert et Kazai comme technique d'évaluation officielle de la première campagne INEX [3].

Ces résultats confirment le fait que les améliorations apportées par les phrases se situent dans les hauts niveaux de rappel. A un niveau de précision élevé, «l'amélioration» est en réalité une dégradation de la performance.

Il est à noter que le très faible résultat de précision@100 pour les SFM seules s'explique par le fait que beaucoup de requêtes contiennent des phrases qui apparaissent dans moins de 100 documents.

6 Perspectives et Conclusions

Nous avons présenté et appliqué un nouveau type de phrases au problème de la recherche d'information documentaire. Nous avons développé et implémenté une technique d'utilisation des séquences fréquentes maximales en recherche d'information. Les expériences menées en utilisant la collection INEX ont donné des résultats encourageants.

Toutefois, l'algorithme d'extraction des SFM est encore lent quand la taille des documents est importante. Cet algorithme a déjà connu de nombreuses améliorations mais il reste certainement perfectible. La technique de partition de la collection en sous-collections permet cependant d'obtenir une approximation en un temps raisonnable.

Il nous faut également expérimenter avec d'autres collections, afin de pouvoir comparer directement nos résultats aux autres. Il est possible que nos résultats soient partiellement dûs à la spécificité de la collection de documents utilisée. Il s'agit en effet d'articles scientifiques, le vocabulaire employé y est donc particulier.

Nos résultats confirment que l'utilisation de phrases améliore les résultats dans les hauts niveaux de rappel. Notre technique est donc certainement plus appropriée à des requêtes d'utilisateurs souhaitant voir la majorité des résultats pertinents. Ce besoin de résultats exhaustifs se trouve par exemple dans le domaine judiciaire, et dans la recherche de brevets.

L'utilisation de phrases est factuelle dans de nombreux langages, ce qui nous rend optimiste quant à de futures expériences avec des corpus multilingues. Le gap doit en outre conférer une robustesse certaine face aux difficultés du multilinguisme.

La découverte des SFM basée sur leur fréquence documentaire reste sans doute un obstacle pour des documents de grande taille. Il serait sans doute opportun de décomposer les articles en sous-documents et de compter le nombre d'occurrences des phrases candidates dans ces sous-documents. Cela donnerait un plus grand nombre de SFM et permettrait d'avoir un poids plus important pour une phrase apparaissant plusieurs fois dans un même document. Différentes granularités sont possible pour définir ces sous-documents et la structure logique de la collection INEX se prête bien à ce type d'expériences.

Références

[1] H. Ahonen-Myka, "Finding All Frequent Maximal Sequences in Text," *16th International Conference on Machine Learning*, pp. 11-17, 1999.

[2] J. L. Fagan, "The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval," *Journal of the American Society for Information Science*, Vol. 40, pp. 115-132, 1989.

[3] N. Gövert, G. Kazai, "Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002," *1st INEX Workshop*, pp. 1-17, 2003.

[4] J. H. Lee, "Combining multiple evidence from different properties of weigthting schemes," *SIGIR*, pp. 180-188, 1995.

[5] M. Mitra, C. Buckley, A. Singhal, C. Cardie, "An analysis of statistical and syntactic phrases," *RIAO97, Computer-Assisted Information Searching on the Internet*, pp. 200-214, 1987.

[6] V. V. Raghavan, P. Bollmann, and G. S. Jung, "A critical investigation of recall and precision as measures of retrieval system performance.," *ACM Transactions on Information Systems*, Vol. 7 (3), pp. 205-229, 1989.

[7] G. Salton, C.S. Yang, C.T. Yu, "A Theory of Term Importance in Automatic Text Analysis," *Journal of the American Society for Information Science*, Vol. 26, pp.33-44, 1975.

[8] A. F. Smeaton, F. Kelledy, "User-chosen phrases in interactive query formulation for information retrieval," *20th BCS-IRSG Colloquium*, 1998

[9] A Turpin, A. Moffat, "Statistical Phrases for Vector-Space Information Retrieval," *SIGIR 1999*, pp. 309-310, 1999.

[10] C. C. Vogt, G.W. Cottrell, "Predicting the performance of linearly combined IR systems," *SIGIR*, pp. 190-196, 1998.

[11] P. Willett, "Recent trends in hierarchic document clustering : a critical review," *Information Processing and Management*, Vol. 24 (5), pp. 577-597, 1988.

[12] Zhai, Chengxiang, Xiang Tong, N. Milic Frayling, D.A. Evans, "Evaluation of Syntactic Phrase Indexing," *TREC-5*, pp. 347-358, 1997.