La reformulation hybride des requêtes exploratoires à l'aide de concepts explicites et implicites

Bissan AUDEH — Philippe BEAUNE — Michel BEIGBEDER

ISCOD, Institut Henri FAYOL
École Nationale Supérieure des Mines de Saint-étienne
158 Cours Fauriel, CS62362, 42023 Saint-Étienne, France
{audeh,beaune,mbeig}@emse.fr

RÉSUMÉ. *Les requêtes exploratoires du Web sont souvent courtes et ambigües. De plus, l'utilisation des entités nommées dans ce type de requêtes est fréquente. Dans cet article, nous explorons l'expansion des requêtes exploratoires par l'ajout de termes appartenant aux concepts de la requête. Nous distinguons deux types de concepts : explicites, correspondants aux entités nommées ayants des références aux concepts de l'ontologie YAGO, et implicites que nous trouvons à l'aide de l'approche LSI (Indexation par la sémantique latente) sur les documents de retour de pertinence. Nous proposons une modélisation hybride qui permet d'intégrer les termes d'expansion extraits de ces deux types de concepts dans la requête. Nos expériences sur une collection issue du Web (WT10G) montrent que notre approche améliore significativement un modèle de base sans expansion, et elle a une meilleur performance qu'une approche classique de retour de pertinence.*

ABSTRACT. *Explorative web queries are often short and ambiguous. In addition, using named entities in this type of queries is well-known and frequent. In this paper, we explore query reformulation for adding terms belonging to query concepts. We distinguish two types of concepts: explicits ones that correspond to named entities having references to concepts in the Yago ontology, and implicit ones that we discover using Latent Semantic Indexing technique on pseudo relevance feedback documents. We propose a hybrid model that integrates terms extracted from these two types of concepts with the original query. Our experiences on a standard web collection (WT10G) show that our approach significantly improves a baseline with no expansion and performs better than a classical expansion approach based on pseudo relevance feedback.*

MOTS-CLÉS : *Reformulation de la requête, entités nommées, retour de pertinence, LSI, Yago*

KEYWORDS: *Query reformulation, named entity, relevance feedback, LSI, Yago.*

CORIA 2014, pp. 273–286, Nancy, 18-21 mars 2014

1. Introduction

La Recherche d'Information (RI) sur le Web est aujourd'hui une pratique fréquente et populaire. Le contenu hétérogène et varié du Web et la diversité des profils utilisateurs posent beaucoup de défis aux systèmes de recherche d'information. L'une des difficultés majeures pour ces systèmes est la formulation des requêtes. Les requêtes des utilisateurs sur le Web peuvent être divisées en trois catégories (Manning *et al.*, 2008) : les requêtes de navigation qui ciblent une page précise, les requêtes transactionnelles comme pour l'achat ou le téléchargement, et les requêtes informationnelles (ou exploratoires). Dans ce papier, nous nous intéressons à ce dernier type de requêtes où l'utilisateur n'exprime pas de façon précise son besoin d'information. Il a été constaté que ces requêtes sont souvent courtes (Jansen *et al.*, 2000) et ne contiennent pas assez de contexte pour résoudre les ambiguïtés, ce qui peut être problématique pour les modèles de RI. Par exemple, la requête « `Baltimore` » (TREC 478) ne précise pas si l'utilisateur s'intéresse à l'histoire de la ville de Baltimore, des informations géographiques sur cette ville ou à un événement lié à cette ville. Plusieurs études s'intéressent à améliorer les résultats de ce genre de requêtes assez fréquent sur le Web. Une possibilité est de modifier la requête originale de l'utilisateur par son expansion. Bien que l'expansion de la requête soit largement abordée en recherche d'information, sa définition et sa catégorisation peuvent varier selon les études. Nous supposons dans cet article que l'expansion de la requête consiste à ajouter des termes à la requête originale et/ou à reformuler la requête en modifiant sa structure et la pondération des termes. Nous précisons qu'un terme est une unité de la requête qui peut être composé d'un ou plusieurs mots.

L'expansion de requête se basant sur la collection de documents est une approche souvent utilisée en recherche d'information. Beaucoup de méthodes dans cette approche utilisent le retour de pertinence aveugle (PRF[1]), qui ne demande pas l'intervention de l'utilisateur. Dans ces méthodes on suppose que les premiers documents rendus par l'itération initiale sont pertinents. Traditionnellement la même collection de documents est utilisée pour récupérer les informations de pertinence en exécutant la requête originale, puis pour évaluer la requête étendue. Plusieurs études récentes ont montré qu'une approche PRF peut être plus efficace si on utilise une collection de documents différente de celle utilisée pour l'interrogation finale (Bendersky *et al.*, 2012 ; Deveaud *et al.*, 2013b ; Egozi *et al.*, 2008 ; ALMasri *et al.*, 2013). En général, les méthodes de retour de pertinence utilisent la requête entière pour la première itération, souvent sous la forme d'un sac de mots. Le fait de prendre en compte le contexte de la requête a l'avantage de résoudre implicitement l'ambigüité des termes, par contre, ces méthodes n'ajoutent que des mots pondérés à la requête, les termes composés de plusieurs mots dans les documents pertinents ne sont donc pas pris en compte explicitement par ces approches. Pourtant, l'importance des groupes nominaux, notamment les entités nommées, est bien connue par des études sur les requêtes longues (Maxwell et Croft, 2013 ; Huston et Croft, 2010 ; Kumaran et Carvalho, 2009). Dans ces études,

1. *Pseudo Relevance Feedback.*

les groupes nominaux sont construits à partir des termes existants dans la requête. Pour les requêtes courtes, cette pratique est moins explorée. Dans cet article, nous proposons un modèle qui permet d'intégrer à une requête des termes d'expansion provenant de deux ressources de nature différente : la collection cible de documents et une ontologie. Bien que le fait de combiner plusieurs ressources pour la reformulation de la requête ne soit pas nouveau (Bendersky *et al.*, 2012 ; Deveaud *et al.*, 2013b), notre contribution est différente par deux aspects : premièrement, les approches précédentes utilisent la même technique d'expansion (le retour de pertinence aveugle) sur des ressources différentes en contenu mais identiques en structure (collections de documents textuels), alors que nous utilisons une collection de documents et une ontologie, ce qui soulève de nouvelles questions sur la gestion de la désambigüisation, la pondération et la reformulation. En deuxième lieu, nous prêtons une attention particulière aux entités nommées, qui peuvent être composées d'un ou plusieurs mots. Bien que la modélisation de (Bendersky *et al.*, 2012) permette de détecter les groupes nominaux (donc éventuellement des entités nommées) à partir des termes de la requête originale, le modèle proposé par les auteurs n'explore pas l'expansion explicite de ces éléments.

La contribution de notre travail est donc de proposer une approche hybride d'expansion des requêtes web exploratoires en nous fondant sur le principe de concepts. Notre idée est de partir de l'hypothèse que chaque terme de la requête cache un concept, puis d'ajouter d'autres termes (expansions) pour chacun de ces concepts. Cette contribution contient trois parties : l'expansion des entités nommées (Section 2), l'expansion des autres termes de la requête par une approche de PRF et des concepts cachés (Section 3), et la modélisation qui permet de fusionner les termes extraits de ces deux parties dans la requête finale (Section 4). Ainsi, à cause de la diversité de ces trois parties, et pour une meilleur lisibilité de l'article, les section 2, 3 et 4 commencent par un panorama sur les travaux proches puis une présentation de notre choix ou de notre proposition. La section 5 présente un exemple démonstratif pour clarifier l'utilisation de l'approche, et la section 6 est consacrée à l'évaluation. Nous terminons l'article par une conclusion et des perspectives en section 7.

2. Les entités nommées comme des concepts explicites

Les entités nommées sont des éléments intéressants dans le domaine du traitement du langage naturel. Dans la littérature de ce domaine on trouve beaucoup d'approches qui s'intéressent à la détection, la classification, ou la désambigüisation des entités nommées. Nous focalisons cette section sur l'utilisation des entités nommées en recherche d'information, notamment pour l'expansion de la requête.

2.1. *Les travaux proches*

En recherche d'information, l'utilisation fréquente des entités nommées dans les requêtes des utilisateurs sur le Web (Guo *et al.*, 2009), a motivé plusieurs études. Ces études proposent des approches qui profitent de l'apparition de ces éléments au niveau

de la requête. Par exemple, (Xu *et al.*, 2008) propose la reformulation de la requête en se basant sur les entités nommées et Wikipedia, et (Wladmir Cardoso Brandao, Altigran Soares da Silva, Edleno Silva de Moura, 2011) proposent une méthode basée sur les pages et les *infobox*[2] de Wikipedia pour trouver des expansions des entités nommées. D'autres travaux prennent en compte les entités nommées pour l'indexation des documents et des requêtes comme (Ngo et Cao, 2011) qui propose une extension du modèle vectoriel basée sur l'indexation par concepts à l'aide de plusieurs ontologies. La méthode que nous proposons s'occupe de la reformulation de la requête en supposant un index classique par racines/mots. Pour cette raison, les approches d'expansion qui nécessitent une indexation spécifique (par exemple conceptuelle) ne rentrent pas dans notre état de l'art.
Bien que beaucoup d'informations sémantiques liées aux entités nommées puissent enrichir le contexte d'une requête, nous focalisons ce papier sur la relation de synonymie. Sous réserve d'une bonne désambigüisation, nous supposons que le choix de la relation de synonymie pour les entités nommées permet de limiter l'effet de dérive de la requête, contrairement aux relations qui généralisent (hyperonymie) ou précisent (hyponymie). Les autres types de relations disponibles dans les ontologies sont souvent différentes selon le type de l'entité, par exemple dans l'ontologie YAGO (Suchanek et Weikum, 2007) on trouve la relation « `est né à` » pour une personne et la relation « `surface` » pour une ville. Il est intéressant de choisir la relation sémantique pour l'expansion d'une façon dynamique selon le type de l'entité nommée et le contexte de la requête, cette piste fera l'objet de nos prochaines études.

2.2. *Notre approche*

Dans cet article, nous sommes dans un contexte générique (Web), et l'utilisation d'une ressource sémantique doit être cohérent avec ce contexte. Bien que WordNet soit une ressource générique qui emploie bien la relation de la synonymie, des études de l'expansion de la requête utilisant cette ressource (Voorhees, 1994) ont rencontré la problématique du manque d'entités nommées dans cette ressource. Pour cela, nous choisissons l'ontologie YAGO qui est construite en utilisant WordNet et les pages des catégories de Wikipédia. Sur cette ontologie, nous appliquons l'approche d'expansion que nous avons proposée dans des études précédentes (Audeh *et al.*, 2014a ; Audeh *et al.*, 2014b) pour l'expansion des entités nommées. Dans cette approche, les entités nommées sont identifiées de la requête par l'outil *Stanford NER*[3]. Chaque entité détectée est désambigüisée grâce aux autres mots dans la requête en utilisant l'implémentation de (Hoffart *et al.*, 2011). Cette implémentation comprend plusieurs techniques pour la désambigüisation : le sens le plus utilisé, la similarité (syntaxe et surface commune) entre une entité nommée et le concept correspondant dans l'ontolo-

2. Une infobox sur Wikipédia est une table préformatée de données dynamiques qui présente sommairement des informations importantes sur un sujet.
3. *Stanford Named Entity Recognizer* (`http://nlp.stanford.edu/software/CRF-NER.shtml`)

Topic	Entité nommée	Variations sémantiques
515	Alexander Graham Bell	« `Alexander Gram Bell` », « `Aleck Bell` », « `The father of the deaf` »
517	Titanic	« `Jinx Titanic` »
478	Baltimore	« `Baltimore City` », « `City of Baltimore` », Baltamore, Bmore, « `Baltimore riots` », Baltimoreans, « `Charm city` », Mobtown, « `Charm City` »

Tableau 1. *Exemples d'appellations trouvés dans YAGO pour des entités nommées.*

gie, et le *graphe de cohérence* entre les concepts. Les deux dernières techniques sont plus utiles avec un texte suffisamment long, ce qui n'est pas le cas pour les requêtes courtes du Web. Dans ces cas, c'est souvent le sens le plus commun qui est choisi pour l'entité nommée par la méthode de (Hoffart *et al.*, 2011). Dans tous les cas, nous obtenons pour chaque entité nommée de la requête une référence unique dans l'ontologie Yago correspondant ainsi à un concept explicite. Les termes d'expansion sont ensuite extraits par la relation « `rdf:Label` » qui donne les appellations, qui peuvent être considérées comme des synonymes, pour l'entité concernée. Le tableau 1 contient quelques exemples d'entités nommées avec leurs expansions possibles. Comme on peut le constater sur ce tableau, les expansions peuvent être des entités nommées ou bien des groupes nominaux qui décrivent l'entité. Par exemple, « `Alexander Graham Bell` » a comme expansions dans Yago « `Alexander Gram Bell` », « `Aleck Bell` » et « `The father of the deaf` ». Après cette étape, nous obtenons pour chaque entité nommée t_e de la requête un ensemble de termes d'expansion $E(t_e)$.

3. L'expansion par des concepts cachés

L'utilisation d'une ontologie pour trouver des termes d'expansion d'entités nommées est efficace car il permet d'obtenir des expressions et des variations qu'on ne peut pas obtenir par les méthodes basées sur le retour de pertinence. Par contre, des informations supplémentaires sur le contexte de la requête sont aussi nécessaires pour connaitre les concepts cachés derrière les autres termes de la requête. Les méthodes de retour de pertinence peuvent être efficaces pour résoudre ce problème.

3.1. *Les travaux proches*

Les méthodes de pseudo retour de pertinence (PRF) sont bien connues en recherche d'information. Certaines de ces méthodes sont complètement dépendantes du modèle de recherche d'information, par exemple le PRF par l'approche de (Rocchio *et al.*, 1965) est adaptée au modèle vectoriel. D'autres approches utilisent des méthodes statistiques sur les documents de retour de pertinence pour extraire des

concepts de façon indépendante du modèle de recherche (Deveaud *et al.*, 2013a). Parmi les nombreuses méthodes d'expansion de la requête par retour de pertinence, nous adoptons une approche inspirée du travail de (Zhao et Callan, 2010) sur l'importance des termes de la requête. Dans leur papier, les auteurs utilisent le retour de pertinence pour créer une matrice termes/documents sur les documents les plus pertinents rendus par la première recherche. Cette matrice est ensuite décomposée en utilisant la technique de décomposition en valeurs singulières (*Singular Value Decomposition*, SVD) qui génère, entre autres, une matrice M_c. Cette dernière matrice donne les composants des termes, considérés comme des vecteurs, dans la base de concepts. Grâce à cette technique, (Zhao et Callan, 2010) calculent plusieurs caractéristiques qui aident à estimer un poids pour chaque terme de la requête sans chercher à trouver des expansions pour ces termes. Nous explorons cette méthode pour trouver des termes sémantiquement proches des termes de la requête. Cela ressemble au travail de (Deveaud *et al.*, 2013a) qui utilise LDA (*Latent Dirichlet Allocation*) au lieu de LSI. La différence est que (Deveaud *et al.*, 2013a) cherche à extraire les concepts cachés de la requête sans forcément faire le lien direct avec chaque terme original, c'est-à-dire que le nombre de concepts est déterminé par une méthode d'apprentissage non supervisé et peut excéder le nombre de termes de la requête. Dans notre approche, nous commençons par supposer que chaque terme de la requête appartient à un concept, et l'algorithme détecte ensuite si certains de ces concepts sont sémantiquement identiques. C'est ce que nous allons voir dans la section suivante.

3.2. *Notre approche*

Contrairement aux travaux précédents, nous explorons LSI pour l'expansion de chaque terme de la requête qui ne fait pas partie d'une entité nommée. Par contre, même si elles ne seront pas étendues par cette approche, les entités nommées ne sont pas exclues du lancement de la requête qui permet d'obtenir les documents de retour de pertinence.

Notre approche commence par la création de la matrice M_c grâce aux documents de retour de pertinence rendus par la requête originale entière. Pour chaque mot dans la requête t_w, qui n'est pas une entité nommée, grâce à M_c, nous pouvons calculer $\cos(\overrightarrow{t}, \overrightarrow{t_w})$, qui donne une similarité entre un terme t et t_w. Nous gardons comme termes d'expansion du terme t_w les n termes qui ont la plus grande similarité, formant ainsi l'ensemble d'expansion $E(t_w)$ correspondant à ce terme ; de plus, cette valeur de similarité nous servira de pondération $s(t, t_w)$ dans la reformulation de la requête (section 4). Nous distinguons deux cas particuliers :

– L'élimination de termes : dans certains cas, des termes de la requête n'apparaissent pas dans les documents de PRF, par conséquent ils n'auraient pas de vecteurs dans la matrice M_c. Dans ce cas, l'ensemble d'expansion pour ces terme sera vide ($E(t_w) = \phi$) et le terme concerné n'apparaîtra pas dans la partie reformulée de la requête finale.

– La fusion des expansions : quand deux termes de la requête sont fortement liés (l'existence de l'un dans un texte signifie l'existence de l'autre également), ces deux termes vont avoir les mêmes statistiques dans les documents de PRF, et on va se retrouver avec les mêmes expansions pour les deux termes. Dans ce cas, nous considérons ces deux termes comme membres du même concept. Pour cette raison, notre algorithme prend en compte les termes de la requête qui n'appartiennent pas aux concepts correspondants aux autres termes déjà traités dans la requête.

4. L'approche d'Expansion Hybride (*EH*) pour la mise en correspondance

Une fois les ensembles d'expansions correspondant aux termes de la requête obtenus, ils doivent être intégrés dans la requête originale. Le choix d'une méthode d'intégration des termes signifie également le choix d'une approche d'évaluation des documents, car le langage des requêtes est fortement lié au modèle de recherche d'information. Dans cet article nous nous intéressons aux modèles de langues bien connus en recherche d'information. Plusieurs études ont proposé des méthodes de recherche conceptuelle basées sur les modèles de langue, où la requête est formulée d'une façon qui permet de prendre en compte les concepts au lieu des termes simples. Par contre, la notion de concepts est différente d'un papier à l'autre. Par exemple, pour (Bendersky et Croft, 2008 ; Metzler et Croft, 2005 ; Maxwell et Croft, 2013), les concepts sont des groupes de termes choisis à l'intérieur des requêtes longues, alors que (Bendersky *et al.*, 2011 ; Bendersky *et al.*, 2012) considèrent deux types de concepts, explicites, c'est à dire composés de termes de la requête, et implicites, c'est-à-dire provenant des documents PRF. D'un autre coté, (Deveaud *et al.*, 2013a) considèrent qu'un concept est l'ensemble de mots pondérés obtenus par LDA (l'allocation latent de Dirichlet) sur les documents de PRF. Dans notre étude, un concept (explicite ou implicite) est un ensemble de termes sémantiquement similaires. Nous rappelons que notre utilisation des termes *explicite* et *implicite* est différente de celle de (Bendersky *et al.*, 2011 ; Bendersky *et al.*, 2012) : les expansions obtenues par les concepts explicites sont les diverses appellations (termes simples ou composés) des entités nommées de la requête, par contre, les expansions provenant des concepts implicites sont composés des mots pondérés qu'on découvre grâce à LSI sur les document de PRF.
Pour intégrer ces termes d'expansion de nature différente, les modèles d'intégration précédemment mentionnés ne sont pas adaptés à notre cas. Ces modèles considèrent le retour de pertinence comme la seule méthode pour obtenir des concepts non composés des termes originaux de la requête, la notion de concepts explicites selon notre vision n'est donc pas modélisée par ces méthodes. L'approche que nous proposons est inspirée par le principe de *Concept Clé* de (Bendersky et Croft, 2008), qui a proposé la reformulation de la requête par l'équation 1 :

$$rank(d) \propto \lambda p(q|d) + (1-\lambda) \sum_{c_i \in q} p(c_i|q) p(c_i|d) \qquad [1]$$

où c_i sont les concepts clés de (Bendersky et Croft, 2008), λ est un paramètre entre 0 et 1, et $p(q|d)$ est la probabilité standard dans les approches de modèle de langue.

Dans notre approche, chaque terme de la requête t_i est remplacé par un ensemble de termes d'expansions $E(t_i)$. Nous remplaçons les concepts clés c_i par les expansions $E(t_i)$ comme dans la formule 2 :

$$rank(d) \propto \lambda p(q|d) + (1-\lambda) \sum_{t_i \in q} p(E(t_i)|q) p(E(t_i)|d) \quad [2]$$

Nous ignorons la probabilité $p(E(t_i)|q)$ car nous supposons que tous les ensembles d'expansion ont la même importance par rapport à la requête. Cette hypothèse forte est justifiée par la section 3.2, où nous avons expliqué que le filtrage des termes non pertinents et la fusions des ensembles d'expansion sémantiquement identiques se font automatiquement par notre approche. Les ensembles d'expansions obtenus par nos approches d'expansion sont donc des concepts pertinents qui ont la même importance pour la requête. L'équation 2 est donc réduite à l'équation 3.

$$rank(d) \propto \lambda p(q|d) + (1-\lambda) \sum_{t_i \in q} p(E(t_i)|d) \quad [3]$$

La probabilité $p(E(t_i)|d)$ est différente selon l'origine de l'ensemble $E(t_i)$. Si le terme t_i est une entité nommée, l'ensemble $E(t_i)$ contient des expansions explicites qui doivent être traités comme équivalents aux termes correspondants dans la requête, c'est-à-dire qu'une occurrence de n'importe quel expansion doit contribuer de manière équivalente au terme original pour le calcul du score d'un document. Ce comportement correspond à l'opérateur $\#syn$ dans le langage de requête d'Indri (Strohman *et al.*, 2004).
Pour les autres termes de la requête qui ne sont pas des entités nommées, $E(t_i)$ contient les termes les plus similaires à t_i dans la matrice M_c. L'équation 4 est utilisée pour calculer la probabilité $p((Et_i)|d)$ pour ces termes.

$$p(E(t_i)|d) = \sum_{w_j \in E(t_i)} p(w_j|E(t_i)) p(w_j|d) \quad [4]$$

La probabilité $p(w_j|E(t_i))$ dans cette équation correspond à la pondération du termes w_j par rapport aux autres termes appartenants au concept exprimé par les termes $E(t_i)$. Nous considérons cette pondération comme la similarité entre le terme d'expansion w_j et le terme de la requête t_i, ce qui est calculé par la cosinus entre les vecteurs de ce deux termes dans la matrice M_c comme expliqué dans la section 3.2. Cette considération est inspiré par le travail de (Ding, 1999) sur le lien entre la similarité et la probabilité dans le cadre de l'analyse sémantique latent (LSI).

5. Exemple démonstratif

Nous prenons dans cette section l'exemple de la requête numéro 455 de TREC 9 (après la suppression des mots vides) :

```
Jackie Robinson appear first game
```

Comme on peut le voir, cette requête contient une entité nommée et trois autres mots. Notre approche va d'abord détecter l'entité nommée « `Jackie Robinson` » qui sera identifiée par une référence unique dans l'ontologie Yago grâce à la méthode de désambigüisation. Cette référence a deux appellations (*Labels*) possibles dans Yago : « `Jackie Robinson` » et « `Jack Roosevelt Robinson` » qui seront considérés comme des expansions dans la requête reformulée. Parallèlement, la requête originale est lancée sur la collection de documents, et l'approche LSI est appliquée sur les documents de PRF. Cette opération donne deux groupes d'expansion pour notre requête démonstrative, chaque mot dans ce groupe est pondéré par son score de similarité avec le terme correspondant de la requête. Avec le langage de la requête de l'outil Indri (Strohman *et al.*, 2004), la modélisation exprimée par l'équation 3 peut être interprétée comme suit :

```
#weight(
   0.8#combine( Jackie Robinson appear first game)
   0.2#combine(
      #syn(#1(Jackie Robinson) #1(Jack Roosevelt Robinson))
      #weight(0.54 girls 0.53 comes
              0.53 uses 0.51 Walsh 0.49 Says )
      #weight(0.82 play 0.81 team 0.79
              games 0.78 season 0.77 players )))
```

Il est intéressant de noter que les trois mots de la requête (« `appear first game` ») n'ont que deux groupes d'expansion dans la requête reformulée, ce qui correspond aux cas particuliers que nous avons expliqués dans la section 3.2. La précision moyenne de cette requête reformulée est améliorée de 24.57% par rapport à la requête originale sous forme d'un sac de mots.

6. Évaluation

La collection de test que nous utilisons est la collection WT10G issue du Web. Les requêtes de nos tests sont les « `titles` » des *topics* TREC 451-550. Vingt-six de ces requêtes contiennent au moins une entité nommée. La collection et les requêtes sont racinisés en utilisant l'approche de Porter et les mots vides ont été enlevés. L'outil de recherche Indri[4] est utilisé pour l'indexation et pour la recherche.

Les différents étapes de notre proposition contiennent plusieurs paramètres qui doivent être réglés. Dans cette étude, les valeurs des paramètres ont été choisies d'une façon expérimentale, mais nous avons conscience de la nécessité d'étudier des alternatives automatiques pour cette étape. Pour LSI, le nombre de documents de PRF et la taille de la matrice réduite ont été fixé à 200 et 150 respectivement en suivant le réglage fait par (Zhao et Callan, 2010) pour la même méthode et sur la même collection. Ces paramètres permettent une utilisation efficients de LSI, qui est connu pour sa complexité de calculs assez élevée avec des grandes matrices. Les valeurs de λ de

4. http ://www.lemurproject.com

Nom	Description
Indri	Requête originale en format sac de mots
Indri-PRF	L'expansion locale de la requête de l'outil Indri
EH-Yago	L'approche d'Expansion Hybride par Yago uniquement
EH-Lsi	L'approche d'Expansion Hybride par LSI uniquement
EH-Hybride	L'approche d'Expansion Hybride par Yago et LSI

Tableau 2. *Notation des approches évaluées.*

l'équation 3 ont été testées dans le domaine $[0{,}1; 0{,}9]$ avec un pas de $0{,}1$, le meilleur MAP a été obtenu pour $\lambda = 0{,}8$ ce qui correspond aux proportions entre la requête sac de mot et la requête conceptuelle de (Bendersky et Croft, 2008) sur la même collection. Le nombre de termes d'expansion dans nos approches pour chaque terme de la requête est 5, que ce soient des expansions explicites ou implicites.

6.1. *Méthodologie*

Le but de nos expériences est d'étudier l'avantage de la reformulation de la requête par des termes provenant de ressources hétérogènes. Pour évaluer l'approche, nous la comparons avec deux méthodes de référence, l'une sans expansion des requêtes, et l'autre avec expansion. Le choix de ces modèles de référence pour évaluer une nouvelle approche est une étape importante. Pour cette étude, notre modèle de recherche de référence (sans expansion) est le modèle par défaut de Indri qui mixe un modèle de langue et un réseau d'inférence[5]. Pour l'approche de PRF nous choisissons l'expansion par approche locale (Ponte et Croft, 1998), fournie par l'outil Indri. Cette approche a obtenu la meilleure performance avec 20 documents de PRF, 10 termes d'expansion et $0{,}5$ comme valeur de λ. Nous gardons donc ces valeurs lors de nos comparaisons. Le tableau 2 donne la notation des approches que nous étudions.

6.2. *Résultats*

Nous avons évalué nos expériences du point de vue de la précision. Ce choix correspond à notre travail avec des requêtes Web, où l'utilisateur n'est pas prêt à chercher des documents pertinents très loin dans la liste de résultats. Pour cela, les métriques que nous avons utilisées sont le MAP (*Mean Average Précision*), la précision à 10 (P10) et la précision à 20 (P20). Les résultats sont présentés dans le tableau 3. Dans ce tableau les meilleurs résultats sont en gras. Les tests de significativité utilisées sont le t-test et le randomization test selon la recommandation de (Smucker *et al.*, 2007). L'étoile (*) signifie une amélioration statistiquement significative (p<0,05) pour ces

5. Nous préparons d'autres comparaisons avec des approches de l'état de l'art pour la suite de cette recherche.

	MAP	P10	P20
Indri	21.47	30.41	27.24
Indri-PRF	**22.60**	30.82	26.58
EH-Yago	21.90	30.61	27.55
EH-Lsi	21.98	32.14*	27.55
EH-Hybride	22.27*	**32.65***	**27.91**

Tableau 3. *Les résultats d'évaluation des approches d'expansion.*

deux tests.
On peut constater dans ce tableau que notre approche EH-Hybride a réussi à améliorer la précision par rapport au modèle de référence sans expansion, cette amélioration est statistiquement significative pour MAP et la précision à 10. Nous constatons également que bien que Indri-PRF améliore mieux le le MAP, cette amélioration n'est pas significative, cette approche dégrade même la précision à 20 du modèle de référence. L'utilisation d'une seule ressource pour l'expansion de la requête améliore légèrement la performance, mais l'intégration des expansions implicites et explicites par notre modélisation *EH-Hybride* est plus significative malgré le petit nombre de requêtes qui contiennent des entités nommées (26 sur 100).

6.3. *Discussion*

Les résultats présentés dans la section précédente montrent l'avantage de combiner plusieurs ressources pour l'expansion de la requête. Bien que ce constat ne soit pas nouveau, notre étude fait le point sur l'intérêt de considérer explicitement les entités nommées. De plus la modélisation que nous proposons respecte la forme de la requête originale formalisée par l'utilisateur. Les approches précédentes fondées sur le retour de pertinence construisent les concepts indépendamment des termes de la requête, alors que nous faisons un lien direct avec chaque terme de la requête. L'avantage de faire ce lien direct entre les termes et leurs expansions est surtout au niveau de l'interprétation : dans un contexte où l'utilisateur peut visualiser la requête reformulée, la requête finale de notre approche peut être facilement comprise par l'utilisateur qui peut l'exécuter telle quelle ou la modifier ou encore l'utiliser comme une base de suggestions, alors que les approches précédentes présentent leurs requêtes étendues comme des sacs de concepts pondérés, où il est plus compliqué de faire le lien avec les termes originaux de la requête. Avec les approches précédentes, le nombre de concepts de la requête est un paramètre qu'il faut considérer, alors que dans notre approche chaque mot dans la requête est considéré comme un concept éventuel qui peut être élargi en ajoutant d'autres termes ou ignoré automatiquement durant la procédure d'expansion.

7. Conclusion

Dans cet article nous avons proposé une nouvelle approche de reformulation de la requête. Cette approche se concentre sur l'intégration des termes d'expansion dans une requête originale en respectant l'hétérogénéité de deux méthodes d'expansion : l'expansion des entités nommées au moyen d'une ontologie et l'expansion par une méthode de retour de pertinence basée sur une analyse sémantique des termes dans les documents PRF. Notre approche est construite pour la recherche sur le Web, c'est-à-dire pour des requêtes courtes et exploratoires, ce qui motive notre idée de faire l'expansion en utilisant des expansions obtenues à partir des concepts de la requête, qui permettent aux utilisateurs de mieux comprendre les liens entres les nouveaux termes ajoutés et les termes originaux de la requête. Les résultats obtenus par notre approche montrent une amélioration significative de la précision, contrairement à une approche qui n'utilisent que le retour de pertinence.

Du côté expérimental, bien que la collection WT10G soit une collection cohérente avec l'évaluation dans un contexte Web, il est important d'évaluer notre approche sur d'autres collections de tests et de les comparer à d'autres approches de l'état de l'art. De l'autre côté, nous souhaitons automatiser le choix des paramètres de nos méthodes d'expansion afin d'optimiser l'utilisation de l'approche. Parallèlement, Yago étant très riche en relations sémantiques, nous explorerons prochainement d'autres relations que la synonymie pour l'expansion des entités nommées.

8. Bibliographie

ALMasri M., Berrut C., Chevallet J.-P., « Wikipedia-based semantic query enrichment », *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval - ESAIR '13*, ACM, 2013.

Audeh B., Beaune P., Beigbeder M., « Exploring Query Reformulation for Named Entity Expansion in Information Retrieval Categories and Subject Descriptors », *to appear in the 29th Symposium On Applied Computing -SAC*, ACM, 2014a.

Audeh B., Beaune P., Beigbeder M., « L'utilisation des entités nommées pour l'expansion sémantique des requêtes Web », *Proceedings of the 14th francophone conference Extraction et Gestion des Connaissances -EGC*, 2014b.

Bendersky M., Croft W. B., « Discovering Key Concepts in Verbose Queries », *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2008.

Bendersky M., Rey M., Croft W. B., « Parameterized Concept Weighting in Verbose Queries », *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2011.

Bendersky M., Rey M., Croft W. B., « Effective Query Formulation with Multiple Information Sources », *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, ACM, 2012.

Deveaud R., Bonnefoy L., Bellot P., « Quantification et identification des concepts implicites d'une requête », *CORIA 2013, La dixième édition de la COnférence en Recherche d'Information et Applications*, UNINE, 2013a.

Deveaud R., SanJuan E., Bellot P., « Estimating topical context by diverging from external resources », *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13*, ACM, 2013b.

Ding C. H. Q., « A Similarity-based Probability Model for Latent Semantic Indexing », *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 1999.

Egozi O., Gabrilovich E., Markovitch S., « Concept-Based Feature Generation and Selection for Information Retrieval », *In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence, 2008.

Guo J., Xu G., Cheng X., Li H., « Named entity recognition in query », *Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2009.

Hoffart J., Yosef M. A., Bordino I., Fürstenau H., Pinkal M., Spaniol M., Taneva B., Thater S., Weikum G., « Robust Disambiguation of Named Entities in Text », *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ACM, 2011.

Huston S., Croft W. B., « Evaluating Verbose Query Processing Techniques », *Proceedings of the 33th international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2010.

Jansen B. J., Spink A., Saracevic T., « Real life, real users, and real needs : a study and analysis of user queries on the web », *Information Processing & Management*, 2000.

Kumaran G., Carvalho V. R., « Reducing long queries using query quality predictors », *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, ACM, 2009.

Manning C. D., Raghavan P., Schütze H., « Web search basics », *Introduction to Information Retrieval*, Cambridge University Press, chapter 19.4, 2008.

Maxwell K. T., Croft W. B., « Compact query term selection using topically related text », *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2013.

Metzler D., Croft W. B., « A Markov random field model for term dependencies », *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05*, ACM, 2005.

Ngo V. M., Cao T. H., « Discovering Latent Concepts and Exploiting Ontological Features for Semantic Text Search », *IJCNLP*, 2011.

Ponte J., Croft W., « A Language Modeling Approach to Information Retrieval », *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 1998.

Rocchio J., Salton J., G, « Information Search Optimization and Iterative Retrieval Techniques », *Fall Joint Computer Conference*, 1965.

Smucker M., Allan J., Carterette B., « A comparison of statistical significance tests for information retrieval evaluation », *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, ACM, 2007.

Strohman T., Metzler D., Turtle H., Croft W., « Indri : A language-model based search engine for complex queries », *Proceedings of the International Conference on Intelligence Analysis,*, 2004.

Suchanek F. M., Weikum G., « YAGO : A Core of Semantic Knowledge Unifying WordNet and Wikipedia », *Proceedings of the 16th international conference on World Wide Web*, ACM, 2007.

Voorhees E. M., « Query expansion using lexical-semantic relations », *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Springer-Verlag New York, Inc., 1994.

Wladmir Cardoso Brandao, Altigran Soares da Silva, Edleno Silva de Moura N. Z., « Exploiting entity semantics for query expansion », *IADIS International conference WWW/Internet*, 2011.

Xu Y., Ding F., Wang B., « Entity-based query reformulation using wikipedia », *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, ACM, 2008.

Zhao L., Callan J., « Term necessity prediction », *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, ACM, 2010.