Expansion de requêtes à base de motifs et de plongements de mots pour améliorer la recherche de microblogs

Meryem Bendella — Mohamed Quafafou

Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

RÉSUMÉ. *Les services sociaux de microblogging jouent un rôle important dans notre société. Twitter est l'une des plateformes de microblogging les plus populaires, utilisées par les internautes pour trouver des informations pertinentes (sujets d'actualité, tendances populaires, informations sur certains internautes, etc.). Dans ce contexte, la recherche d'information provenant de telles données a récemment gagné un intérêt majeur et ouvert de nouveaux défis. Cependant, la taille de ces données ainsi que des requêtes est généralement courte et peut avoir un impact sur le résultat de la recherche. Cette dernière peut être améliorée à l'aide de l'expansion de requêtes. En effet, les mots peuvent avoir plusieurs sens dont un seul est utilisé pour un contexte donné. Dans cet article, nous proposons une méthode d'expansion de requêtes prenant en compte le sens du contexte. Nous utilisons les motifs et les plongements de mots pour étendre les requêtes des utilisateurs. L'évaluation expérimentale de la méthode proposée est menée sur la collection TREC. Les résultats montrent l'efficacité de l'approche en combinant des motifs avec des plongements de mots pour améliorer significativement la recherche de microblogs.*

ABSTRACT. *Social microblogging services have an especially significant role in our society. Twitter is one of the most popular microblogging sites used by people to find relevant information (e.g., breaking news, popular trends, information about people of interest, etc). In this context, retrieving information from such data has recently gained growing attention and opening new challenges. However, the size of such data and queries is usually short and may impact the search result. Query Expansion (QE) has a main task in this issue. In fact, words can have different meanings where only one is used for a given context. In this paper, we propose a QE method by considering the meaning of the context. Thus, we use patterns and Word Embeddings to expand users' queries. We experiment and evaluate the proposed method on the TREC dataset. Results show the effectiveness of the proposed approach and signify the combination of patterns and word embedding for enhanced microblog retrieval.*

MOTS-CLÉS : *Expansion de requêtes, Motifs, Plongements de mots, Recherche de microblogs*

KEYWORDS: *Query expansion, Patterns, Word Embeddings, Microblog retrieval*

1. Introduction

Twitter est l'une des plateformes de médias sociaux les plus populaires permettant aux utilisateurs de publier des textes courts (jusqu'à 280 caractères pour un texte) appelés tweets. De plus en plus d'utilisateurs accèdent à ces plateformes pour partager des idées, des opinions, des suggestions, des histoires quotidiennes et des événements. Parmi ces services en ligne, plusieurs commencent à faire partie de la vie quotidienne de millions de personnes dans le monde. Cependant, une énorme quantité d'informations est créée dans ces plateformes, il est donc difficile de trouver des informations récentes et pertinentes.

De nombreux utilisateurs sont intéressés par la collecte d'informations récentes à partir de telles plateformes. Ces informations peuvent être liées à un événement particulier, à un sujet spécifique ou aux tendances populaires. Les utilisateurs expriment leur besoin par une requête pour la recherche des messages publiés (tweets) sur ces plateformes (Twitter). Selon (Spink *et al.*, 2001), la plupart des gens utilisent peu de termes de recherche et peu de requêtes modifiées dans la recherche sur le Web. De ce fait, le système ne retrouve pas ou peu de documents pertinents à la requête de l'utilisateur. Cependant, la formulation de la requête devient difficile pour l'utilisateur afin d'exprimer correctement son besoin. L'expansion de requêtes apporte une contribution significative pour améliorer la qualité et les résultats de recherche dans ce cas. Une requête étendue contiendra des termes plus connexes (appelés termes candidats) afin d'augmenter les chances de retrouver le nombre maximal de documents pertinents. L'objectif est de trouver les microblogs répondant à un besoin d'information spécifié par un utilisateur.

Dans cet article, nous proposons une nouvelle méthode basée sur les motifs fréquents fermés et les plongements de mots (*Word Embeddings*) pour étendre les requêtes des utilisateurs. Pour atteindre cet objectif, nous préparons d'abords notre collection de données en effectuant des pré-traitements du texte des tweets, qui est une étape cruciale dans le domaine de la recherche d'information (RI) et du traitement automatique des langues (TAL). Ensuite, chaque tweet est représenté par un ensemble de mots et sera indexé par le système Terrier. Par la suite, nous utilisons ce système pour retrouver les tweets en fonction des requêtes définies par TREC2011 (Ounis *et al.*, 2011). Nous avons utilisé le modèle BM25 (Jones *et al.*, 2000) pour sélectionner les tweets pertinents répondant aux requêtes initiales. Ces tweets renvoyés par le système sont utilisés pour extraire des motifs. Ceci peut être défini comme des motifs fermés fréquents contenus dans la collection des tweets. Pour l'entraînement des *embeddings* sur notre corpus de données textuelles, nous avons utilisé le modèle Word2Vec (Mikolov *et al.*, 2013). En effet, l'expansion des requêtes initiales est réalisé en combinant les motifs et les plongements de mots. Cette combinaison consiste à enrichir la requête en sélectionnant les mots les plus proches aux mots constituant les motifs. La méthode proposée étend la requête initiale et améliore les résultats de la recherche de microblogs (tweets).

Cet article est organisé comme suit : la section 2 correspond à un un état de l'art des différentes méthodes et travaux d'expansion de requêtes dans la recherche de microblogs. La section 3 décrit toutes les étapes nécessaires de notre approche proposée. La section 4 présente notre méthode d'expansion de requêtes. Les expérimentations ainsi que les résultats obtenus sont présentés dans la section 5. Enfin, la section 6 conclut cet article en donnant quelques perspectives.

2. État de l'art

L'attention accordée à l'expansion de requêtes a récemment augmenté dans le domaine de la RI. De nombreux travaux de recherche ont adressé le problème des requêtes courtes. Les requêtes formulées par les utilisateurs peuvent être courtes et les résultats de la recherche ne sont donc pas centrés sur le sujet d'intérêt.

D'importants efforts ont été consentis pour améliorer la recherche de microblogs. Dans (Massoudi *et al.*, 2011), les auteurs proposent un modèle de recherche pour retrouver des messages de microblog liés à un sujet d'intérêt donné. Ce modèle est fondé sur une combinaison d'indicateurs de qualité et du modèle d'expansion des requêtes. Dans (Anagnostopoulos *et al.*, 2012), les auteurs ont utilisé les hashtags et introduisent une approche permettant de créer un réseau sémantique pouvant être exploité pour permettre l'expansion de requêtes en fonction des besoins d'information des utilisateurs. Dans (Li et Jones, 2017), trois techniques différentes d'expansion de requête ont été étudiées afin d'améliorer les résultats de la recherche. La première technique est basée sur une ressource externe, la deuxième est basée sur WordNet et la troisième utilise une intervention manuelle.

Des méthodes d'expansion basées sur les plongements de mots ont été utilisées dans la littérature. Ils ont été utilisés pour re-pondérer les mots de la requête (Zheng et Callan, 2015) et pour comparer directement une requête à un document (Vulić et Moens, 2015 ; Nalisnick *et al.*, 2016). Dans (Kuzi *et al.*, 2016), les auteurs proposent une méthode basée sur les plongements de mots en sélectionnant les termes qui sont sémantiquement liés à la requête. Ils ont montré que les performances des résultats de recherche utilisant les plongements de mots sont considérablement meilleures que ceux utilisant uniquement la requête initiale.

Les approches d'expansion de requête pour la recherche de microblogs peuvent être divisées en trois groupes, à savoir local, global et externe (Pal *et al.*, 2015) :

– **Local** : Les techniques d'expansion de requêtes locales sélectionnent les termes candidats depuis un ensemble de documents retrouvés en réponse à la requête originale (non étendue). Ce type d'approche est connu sous le nom de *Pseudo Retour de Pertinence* (PRF). Cette dernière est largement utilisée pour l'expansion de requêtes dans la recherche de microblogs (Diaz *et al.*, 2016 ; Massoudi *et al.*, 2011 ; Zhai et Lafferty, 2001). Cette approche consiste à utiliser des termes dérivés des N premiers documents retrouvés (documents pertinents) pour extraire d'autres documents similaires qui sont également susceptibles d'être pertinents. Dans, (Lau *et al.*, 2011), les

auteurs proposent un algorithme d'extraction de traits thématiques à partir des tweets en utilisant les techniques de fouille de motifs. Leur méthode d'expansion de requêtes est basée sur l'approche PRF en appliquant des poids pour les différents traits. Toutefois, dans le cas d'une recherche de tweets, où les requêtes sont courtes, la liste des premiers documents retrouvés peut ne contenir que quelques documents pertinents, ce qui peut entraîner la présence de quelques termes liés à la requête originale (termes candidats). Le travail de cet article aborde ce problème en utilisant des techniques de fouille de données dans lesquelles nous extrayons des motifs importants depuis les N premiers documents pertinents.

– **Global** : Les techniques globales sélectionnent les termes d'expansion à partir de l'ensemble de la base de données documentaire. Ces techniques sélectionnent les termes candidats en extrayant les relations terme-terme à partir du corpus cible (Pal *et al.*, 2015). Ce type de techniques a été utilisé dans de nombreuses applications (Hu *et al.*, 2006 ; Mittal *et al.*, 2010). Il s'agit de l'une des premières techniques à produire des améliorations d'efficacité cohérentes grâce à une expansion automatique (Carpineto et Romano, 2012). Dans (Bai *et al.*, 2005), l'approche proposée consistait à utiliser des relations de terme dans l'expansion de requêtes dans le cadre du modèle de langue (LM). Ils ont également intégré des relations calculées par flux d'informations dans un LM. Les auteurs ont montré que l'idée d'expansion avec des relations entre termes peut être naturellement appliquée dans le modèle de langue. Dans notre travail, nous utilisons d'une certaine manière l'analyse globale en effectuant l'entraînement des plongements de mots sur l'ensemble de données afin d'extraire les termes les plus similaires aux termes des motifs. Dans (Yang *et al.*, 2017), les auteurs proposent une méthode d'expansion de requêtes utilisant les plongements de mots et quelques ressources externes.

– **Externe** : Les techniques externes constituent les méthodes qui permettent d'obtenir les termes d'expansion à partir d'autres ressources en dehors du corpus cible (Pal *et al.*, 2015). Plusieurs approches se concentrent sur l'utilisation des ressources externes telles que *Wikipedia*, *WordNet* et *DBpedia* , afin d'améliorer l'expansion de requêtes (Kotov et Zhai, 2012 ; Aggarwal et Buitelaar, 2012 ; Zingla *et al.*, 2016). Les auteurs dans (Zingla *et al.*, 2016) ont proposé des approches basées sur la technique d'extraction des règles d'association et des sources externes pour étendre les requêtes courtes. Dans (Zingla *et al.*, 2018), les auteurs présentent un modèle d'expansion de requête hybride (HQE) qui étudie la manière dont les ressources externes peuvent être combinées avec l'extraction de règles d'association et utilisées pour améliorer la génération et la sélection de termes d'expansion. En outre, dans (Aggarwal et Buitelaar, 2012) les auteurs proposent une approche pour l'expansion de requêtes utilisant des sources externes. Leur méthode inclut la génération de concepts candidats à partir de Wikipédia et de DBpedia, ainsi que la sélection des K-meilleurs concepts selon les scores de l'analyse sémantique explicite.

3. Expansion de requêtes basée sur l'extraction de motifs

Dans cette section, nous décrivons toutes les étapes requises dans notre approche proposée pour étendre les requêtes dans la recherche de microblogs. Dans un premier temps, nous nous sommes concentrés sur le pré-traitement des données. Ensuite, nous décrivons la tâche d'extraction de motifs et définissons les notions nécessaires à la compréhension de l'approche proposée.

3.1. *Pré-traitement*

Ce processus comprend le prétraitement du texte du tweet, d'où l'élimination des mots outils (*stop words*), des émoticônes et de la ponctuation, la racinisation, etc. Les données de microblogs telles que les tweets sont très courtes, généralement mal écrites et ne respectent pas la grammaire. Cependant, cette étape préliminaire est très cruciale pour éliminer le bruit et nettoyer les données. Nous préparons donc l'ensemble de données pour l'indexation en filtrant les tweets comme suit :

– Suppression des tweets nuls (sans contenu) et des tweets courts contenant moins de deux mots.

– Suppression des retweets (tweets commençant par RT suivis d'un nom d'utilisateur), car ils seraient jugés non pertinents.

– Suppression des tweets qui ne sont pas en anglais.

– Élimination du contenu non-ASCII trouvé dans les tweets en anglais.

– Suppression des liens et des mentions d'utilisateurs des tweets.

– Extraction des hashtags du tweet.

Plus précisément, nous conservons les hashtags (mots précédés par le symbole "#") car ce type d'informations est important sur Twitter. Dans (Anagnostopoulos *et al.*, 2012), les auteurs proposent une approche basée sur les hashtags pour étendre les requêtes dans la recherche de microblogs.

Par ailleurs, tous les mots du tweet sont racinisés sauf les hashtags. Cette étape est effectuée en utilisant la méthode de racinisation de Porter implémenté dans l'outil Stanford NLP[1]. Les mots outils[2] sont supprimés du tweet. Nous effectuons également une normalisation du contenu du tweet, traitant les mots contenant de nombreuses lettres répétées, tels que « yes » ou « happy », qui peuvent apparaître comme « yeeeees » ou « happyyy » sur Twitter. Ces étapes de pré-traitements sont effectuées avant la phase d'indexation des documents (tweets) par le système de RI.

1. https ://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
2. https ://github.com/ravikiranj/twitter-sentiment-analyzer/blob/master/data/feature_list/stopwords.txt

3.2. *Découverte de motifs et Analyse Formelle de Concepts*

La découverte de motifs fréquents est devenue une tâche très importante dans le domaine de fouille de données et dans d'autres domaines d'applications. Elle a été initiée par Agrawal et al. (Agrawal *et al.*, 1993) et elle correspond à trouver les ensembles d'attributs (ou items) qui apparaissent simultanément dans au moins un certain nombre d'objets (ou transactions) définis dans un contexte d'extraction (voir définition 1).

Ensemble de transactions

L'ensemble de transactions est représenté par un ensemble de tweets prétraités. Chaque tweet est représenté par un ensemble de mots considérés comme des items. Le tableau 3.2 donne un exemple de transactions.

Id	**Items**				
t1	A		C	D	
t2		B	C		E
t3	A	B	C		E
t4		B	C		E
t5	A	B	C		E

Tableau 1 : Exemple de transactions

Nous introduisons les définitions suivantes pour l'extraction de motifs :

Définition 1. *(Contexte d'extraction). Un contexte d'extraction $\mathcal{D} = (\mathcal{T}, \mathcal{I}, \mathcal{R})$ est composé d'un ensemble de transactions $\mathcal{T}$, d'un ensemble d'items $\mathcal{I}$ et d'une relation binaire entre les transactions et les items $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{I}$. Chaque couple $(t, i) \in \mathcal{R}$, signifie que la transaction t est en relation avec l'item i.*

Définition 2. *(Motif, Couverture et Support). Un motif X est un sous-ensemble d'items $X \subseteq \mathcal{I}$. Sa couverture et son support sont définis par :*

$$cover(X) = \{t \in \mathcal{T} | \forall i \in X, (t, i) \in \mathcal{R}\} \qquad support(X) = |cover(X)|$$

Définition 3. *(Motif fréquent) Un motif est fréquent si sa fréquence dépasse un seuil minsup fixé par l'utilisateur. Étant donné un contexte d'extraction $\mathcal{D}$ et un minsup (le support minimum), l'ensemble de tous les motifs fréquents dans $\mathcal{D}$ est :*

$$\mathcal{FI} = \{X \subseteq \mathcal{I} | support(X) \geq minsup\}.$$

Dans ce travail, nous nous intéressons aux motifs fermés fréquents. Le motif X est dit fermé si aucun de ses sur-ensembles n'a le même support que X(voir définition 4). D'un point de vue plus pratique, un motif X est fermé s'il est égal à sa fermeture $Closure(X)$ (voir définition 5). Le motif fermé fréquent est un motif fréquent et fermé. Notons qu'il suffit donc de connaître l'ensemble de motifs fermés fréquents (noté $\mathcal{FCI}$) pour pouvoir générer tous les motifs fréquents et leurs supports. Nous avons aussi, $\mathcal{FCI} \subseteq \mathcal{FI}$

Définition 4. *(Motifs fermés) X est fermé $\Leftrightarrow X \subset Y$, $support(Y) < support(X)$*

Définition 5. *(Motif fermé et fermeture) X est fermé $\Leftrightarrow X = Closure(X)$ où*

$$Closure(X) = \bigcap_{t | X \subseteq t} \{t\}$$

Analyse Formelle de Concepts (AFC)

L'analyse formelle de concept (AFC) est une théorie de l'analyse de données qui identifie les structures conceptuelles dans un ensemble de données. Il s'agit d'une méthode de regroupement conceptuel qui fournit une approche menant à la découverte et la structuration de connaissances. Elle présente un formalisme pour identifier les dépendances entre les données (Codocedo *et al.*, 2016). il y a un ensemble ordonné unique qui décrit la structure inhérente en treillis définissant les regroupements naturels et les relations entre les transactions et leurs items associés. Cette structure est connue sous le nom de treillis de concept ou treillis de Galois (Ganter et Wille, 2012). Chaque élément du treillis est un couple (T, I) composé d'un ensemble de transactions (*l'extension*) et un ensemble d'items (*l'intention*). Chaque couple (appelé *concept formel*), doit être un couple complet pour $\mathcal{R}$, ce qui signifie que les applications f et g présentées dans la définition 6 s'appliquent. $f(T)$ renvoie les items communs à toutes les transactions $t \in T$, tandis que $g(I)$ renvoie les transactions qui ont au moins tous les items $i \in I$.

Définition 6. *Pour $T \subseteq \mathcal{T}$ et $I \subseteq \mathcal{I}$, nous avons :*

$$f(T) = \{i \in \mathcal{I} | \forall t \in T, (t, i) \in \mathcal{R}\} \quad et \quad g(I) = \{t \in \mathcal{T} | \forall i \in I, (t, i) \in \mathcal{R}\}.$$

Figure 1 montre le treillis de concept construit à partir du Tableau 3.2. Chaque noeud dans ce treillis correspond à un concept formel, où les noeuds t1, t2, t3, t4, et t5 représentent les objets (transactions ou tweets) et les noeuds A, B, C, D, et E représentent les attributs (items) utilisés pour nommer le concept. Par exemple, le noeud nommé par **B** représente le concept formel ({t2, t3, t4, t5}, {B C}). Le noeud étiqueté par **t5** correspond au concept formel ({t3, t5}, {A B C}).

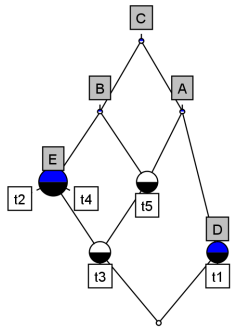

Figure 1 : Exemple d'un treillis de concept

L'idée d'étendre au maximum les ensembles est formalisée par la notion mathématique de *fermeture* dans des ensembles ordonnés. Les opérateurs $h_1 = f \circ g$ et $h_2 = g \circ f$ sont les opérateurs de fermeture de Galois. Soit X un motif d'attributs (items), si $h_1(X)$=X, alors X est un *motif fermé*. Nous pouvons remarquer qu'un concept formel est composé d'un motif fermé et de l'ensemble des transactions contenant ce motif. Dans notre cas, le concept formel est composé d'un ensemble de termes (motif fermé) et de l'ensemble des tweets contenant ce motif.

4. Description de la méthode

Dans cette section, nous décrivons notre approche proposée pour étendre les requêtes dans la recherche de microblogs. La méthode proposée est basée sur l'extraction de concepts fréquents et les plongements de mots. Elle se compose de trois étapes principales : (1) Extraction et sélection des motifs fermés fréquents, (2) Etendre les motifs en utilisant les plongements de mots, et (3) Expansion de requêtes (voir Figure 2).

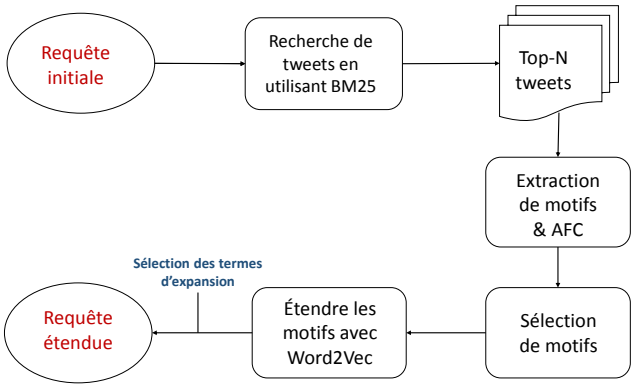

Figure 2 : Vue d'ensemble de l'approche proposée pour l'expansion de la requête.

4.1. *Extraction de motifs fermés fréquents*

Le treillis de concept fréquent est calculé sur les N premiers tweets renvoyés par le système Terrier en réponse aux requêtes initiales. Nous rappelons que le concept est composé d'un motif fermé et de l'ensemble des transactions contenant ce motif. En effet, la découverte de motifs fermés fréquents est réalisée sur N tweets renvoyés par le système. Ce processus est effectué à l'aide de l'algorithme Charm-L (Zaki et Hsiao, 2005), où la découverte de motifs calcule les ensembles d'items fermés (c'est-à-dire les mots) qui apparaissent ensemble dans au moins un certain nombre de transactions (c'est-à-dire des tweets) contenus dans le corpus de données. Ce nombre est appelé *minsup*. Plus précisément, un treillis de concept fréquent est formé en utilisant les concepts formels ayant au moins *minsup* transactions dans leur extension. Chaque nœud du concept formel $\mathcal{L}$ représente la correspondance entre un motif et l'ensemble de tweets contenant tous les mots de ce motif.

L'algorithme permettant de générer l'ensemble de motifs fermés fréquents intéressants est présenté dans Algorithme 1 , avec $\mathcal{L}_K$ désignant l'ensemble des K items

fermés fréquents, T l'ensemble des n transactions qui représentent les tweets, *minsup* la valeur du support minimum (seuil minimum).

Algorithm 1 Extraction de motifs

Entrées :

$\mathcal{Q}$: requête initiale
$\mathcal{T}$: l'ensemble de n tweets correspondant à $\mathcal{Q}$
$\mathcal{W}$: Vocabulaire de tous les mots contenus dans le corpus de tweets
$minsup$: support minimum (seuil minimum)
$\mathcal{R}$: une relation binaire où $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{W}$

Résultat :

$\mathcal{L}_K$: Liste de K motifs

1: Création du contexte formel $\mathcal{D}$ =($\mathcal{T}, \mathcal{W}, \mathcal{R}$) ;
2: Calcul de treillis de concept fréquent $\mathcal{L}$ pour $\mathcal{D}$ en fonction de $minsup$;
3: $\mathcal{L}_K = CharmL(\mathcal{T}$, minsup) ;
4: **Retourner** $\mathcal{L}_K$;

A partir de la liste de K concepts fréquents (motifs et transactions) retrouvés par l'application de l'algorithme 1, nous sélectionnons k motifs afin de représenter les termes candidats pour l'expansion de requêtes. Cette sélection est basée sur le support de chaque motif. La liste est triée par ordre croissant selon la valeur du support (nombre de transactions dans le concept) et les k premiers motifs sont sélectionnés. Par exemple, pour une requête Q_i, nous avons $\mathcal{L}_{K_i}$ motifs. Soit $\mathcal{Q}_P$ l'ensemble de termes candidats. Cet ensemble est obtenu à partir de k motifs où $\mathcal{L}_{K_i} = argmax(Support(\mathcal{L}_{K_i}))$. Nous détaillons dans la section qui suit comment ces motifs sont étendus par les plongements de mots.

4.2. *Plongements de mots : modèle Word2Vec*

Les plongements de mots ont déjà été appliquées avec succès à l'expansion de requêtes pour améliorer la RI (Roy *et al.*, 2016). Dans cet article, nous les combinons avec les motifs pour étendre les requêtes dans la recherche de microblogs (tweets). L'entraînement du réseau de neurones est réalisé sur les mêmes données du corpus de tweets (TREC 2011) afin d'extraire les termes les plus similaires aux motifs sélectionnés calculés dans la phase précédente. Ces mots sont susceptibles d'être pertinents pour la requête mais n'apparaissent pas dans les motifs. L'intérêt d'utiliser les plongements de mots est d'enrichir les termes des motifs par des mots pertinents qui n'ont pas été détectés par l'algorithme d'extraction de motifs. Ces mots pertinents proviennent du corpus analysé (corpus contenant tous les tweets) tandis que les mots présents dans les motifs proviennent des N premiers résultats (tweets) d'une recherche.

Pour l'entraînement des *embeddings*, nous utilisons une alternative du projet WORD2VEC [3] fait en JAVA par l'équipe MEDALLIA[4] pour l'intégrer dans notre programme principal, implémenté en JAVA. Ce modèle estime la probabilité qu'un terme apparaisse à une position dans le texte en se basant sur les termes apparaissant dans une fenêtre autour de cette position. Chaque mot est représenté par un vecteur plein, de taille modérée, qui correspond à une projection du mot dans un espace vectoriel. Les similarités entre ces vecteurs correspondent aux similarités sémantiques entre les mots (termes) (Mikolov *et al.*, 2013).

Pour le paramétrage du réseau de neurones, nous avons choisi une fenêtre d'une taille de 7 mots (leur fréquence d'apparition est d'un minimum égal à 5, la dimension des vecteurs est de 200, le nombre d'exemples négatifs est de 7). Plus précisément, le modèle de sac de mots continu (CBOW) est utilisé.

4.3. *Expansion de requêtes*

Étant donné une requête originale $q = \{w_{q1}, ..., w_{qn}\}$, le processus d'expansion de q a trois objectifs : (1) recherche de tweets pertinents à q à l'aide du modèle BM25, (2) sélection des termes candidats pour q (w_{eq}) en calculant le treillis de concepts fréquents (motifs fermés fréquents) à partir des N premiers tweets renvoyés, (3) sélection des termes les plus liés aux termes candidats en utilisant les plongements de mots, afin d'ajouter uniquement les termes liés sémantiquement à q. Ces termes sont ensuite sélectionnés pour obtenir l'ensemble des termes qui représentent la requête étendue notée eq, avec $eq = q \bigcup \{w_{eq_1}, ..., w_{eq_m}\}$.

Le processus d'obtention des termes candidats consiste à sélectionner les termes à partir des motifs fermés fréquents obtenus en appliquant la méthode proposée dans la section 4.1. Pour enrichir ces termes, nous avons utilisé le modèle *Word2Vec* afin de sélectionner des termes liés sémantiquement aux motifs obtenus. Le processus de sélection de ces termes calcule la similarité entre le vecteur des termes du motif correspondant et chaque vecteur de termes du tweet. Pour mesurer cette similarité, nous utilisons la mesure *cosinus*, les mots sont ensuite triés dans l'ordre décroissant. Le nombre de termes d'expansion n'est pas fixé, il dépend d'une part, de nombre de termes du motif et d'autre part, des *k* termes les plus similaires retrouvés par la mesure de similarité *cosinus*. Le calcul de similarité a été effectué en se basant sur le réseau de neurone entraîné par le modèle *Word2Vec*.

5. Expérimentations

Dans cette section, nous menons des expérimentations pour évaluer l'efficacité de la méthode proposée pour l'expansion des requêtes. Afin de mesurer la performance

3. https://github.com/medallia/Word2VecJava
4. http://engineering.medallia.com

de notre méthode proposée, nous comparons notre méthode d'expansion de requêtes basée sur les motifs avec une *baseline* ainsi que d'autres méthodes proposées dans la littérature. Nos expérimentations sont menées sur la collection TREC 2011 contenant environ 16 millions tweets et 49 requêtes. Nous utilisons les mesures *Précision*, *MAP* et *nDCG* pour évaluer la méthode proposée pour l'expansion de requêtes dans la recherche de microblogs. Ainsi, nous évaluons la performance de la méthode proposée selon les mesures F-mesure et R-PREC afin de comparer notre approche avec quelques approches de la littérature évaluées sur la même collection de données.

5.1. *Description des données*

Nous avons évalué notre méthode sur la collection de test, issue de la tâche Microblog de TREC[5] 2011. Cette collection de données contient environ 16 millions de tweets collectés sur une période de 2 semaines (du 24 Janvier au 8 février 2011) (Ounis *et al.*, 2011). Vu que l'ensemble de données fourni ne contient que des identifiants de tweets, nous avons collecté environ 12 millions de tweets avec du contenu. Quelques tweets ont été supprimés par leurs éditeurs et d'autres nous ne pouvons plus y avoir accès. Après l'application du processus de pré-traitements de données détaillé dans la section 3.1, nous avons obtenu un corpus de données d'environ 3,5 millions de tweets sur lequel nos expérimentations ont été menées.

Nous avons utilisé 49 requêtes définies par TREC 2011 (Ounis *et al.*, 2011). Un exemple d'une requête avec le format TREC est comme suit :

```
<top>
  <num> Number : MB001 </num>
  <title> BBC World Service staff cuts </title>
  <querytime> Tue Feb 08 12 :30 :27 +0000 2011 </querytime>
  <querytweettime> 34952194402811904 </querytweettime>
</top>
```

5.2. *Modèle de recherche*

Nous utilisons le système Terrier[6] qui est une plateforme de RI, pour indéxer notre collection de données (TREC 2011). Cette plateforme offre des fonctionnalités d'indexation et de recherche, des modèles de pondération de documents et d'expansion de requêtes. Terrier a été utilisé avec succès pour différents types de recherche (Ounis *et al.*, 2005). Tous les tweets sont indexés à l'aide de Terrier et les requêtes originales (seront détaillées dans la section 5) sont utilisées pour retrouver et trier les tweets à l'aide du modèle de RI standard BM25 (Jones *et al.*, 2000) de Terrier. Le modèle BM25 a été largement utilisé dans la communauté TREC sur une variété de corpus.

5. Text REtrieval Conference (TREC)
6. `http://terrier.org/`

5.3. *Résultats expérimentaux*

Nous avons effectué deux implémentations différentes dans nos expérimentations que nous avons menées, la première est basée sur des motifs fermés fréquents (nommée Run-P) et la seconde représente la combinaison de ces motifs avec les plongements de mots (nommée Run-P-WE). Le Tableau 5.3 indique les résultats en terme de Précision, MAP et nDCG obtenus par chacune des variantes. Dans ce travail, nous nous sommes basés sur les résultats de recherche obtenus avec le modèle BM25 implémenté dans Terrier comme *baseline*. Ces résultats de recherche constituent les 1000 tweets renvoyés par Terrier en réponse aux requêtes initiales, autrement dit, sans expansion de requêtes. Nous avons également utilisé une méthode basée sur le Pseudo Retour de Pertinence (PRF) afin de comparer nos résultats. Cette méthode d'expansion est appliquée en utilisant le modèle de pondération de termes *Bo1* implémenté dans terrier. Nous avons pris en considération la configuration par défaut proposée par le système (à savoir, le nombre de N premiers documents : 3 ; nombre de termes pour étendre une requête : 10).

Mesures	**P@10**	**P@30**	**MAP**	**nDCG@10**	**nDCG**
Baseline	0.1184	0.0905	0.1025	0.1146	0.2659
PRF	0.2245	0.2116	0.1759	0.2067	0.3876
Run-P	0.2980	0.2415	0.1929	0.2619	0.3938
Run-P-WE	**0.3449**	**0.2878**	**0.2403**	**0.3077**	**0.4476**

Tableau 2 : Résultats d'évaluation de l'approche proposée

L'exploitation de l'analyse formelle de concepts et l'extraction de motifs s'est avérée efficace pour la plupart des requêtes. Néanmoins, en raison de la concision des requêtes ainsi que des microblogs , certaines requêtes donnent une performance médiocre. Pour cela, nous avons utilisé les plongements de mots pour étendre la sémantique du motif et enrichir la requête pour mieux exprimer les besoins d'information. Cette combinaison des deux approches du domaine de fouille de données et du TAL améliorent significativement les résultats d'expansion de requêtes par rapport au modèle *Run-P* et la *baseline*, comme le montre la Figure 3. Le score de précision le plus élevé dans *Run-P-WE* est de 1.00 pour quelques requêtes. Dans l'ensemble, les résultats expérimentaux en appliquant l'approche utilisant des motifs et des plongements de mots surpassent les résultats de l'application de l'approche utilisant uniquement les motifs.

Par rapport à la *baseline*, nous avons obtenu des améliorations sur les quatre mesures : +191,30%, +218,01%, +134,43% et +168,5% respectivement sur la P@10, la P@30, la MAP et le nDCG@10.

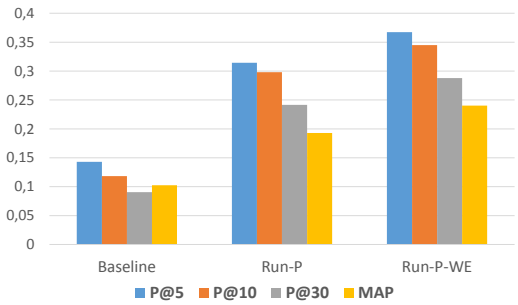

Figure 3 : Comparaison des résultats obtenus avec et sans expansion de requêtes.

Le Tableau 3 montre les résultats de précision en considérant les 30 premiers tweets renvoyés (P@30), de la MAP et de R-Prec (*Reciprocal Precision*) obtenus par l'application de notre approche par rapport aux différentes variantes de l'approche proposée par les auteurs dans (Lau *et al.*, 2011). Le choix de comparaison s'est porté sur cette approche car les auteurs ont utilisé le concept de co-occurrence (motifs fréquents) et ont évalué l'efficacité de l'approche proposée sur la collection de test TREC 2011. Nous avons obtenu des améliorations significatives sur les trois mesures : 41,49%, 219,12% et 107,8% respectivement sur la P@30, la MAP et la R-Prec (voir tableau 3). RUN1 désignant un run en utilisant la fréquence des termes sans pondération. RUN2 est une deuxième configuration où la pondération des termes est basée sur l'occurrence d'un terme dans les motifs. RUN3 est une troisième configuration basée sur les motifs uniquement où la pondération d'un motif s'effectue en utilisant les poids des termes contenus dans le motif. Enfin, RUN4 est la combinaison de la deuxième et de la troisième configuration (Lau *et al.*, 2011).

Mesure/RunID	**Run1A**	**Run2A**	**Run3A**	**Run4A**	**Run-P-WE**
P@30	0.1347	0.1694	0.2034	0.1973	**0.2878**
MAP	0.0753	0.0486	0.0673	0.0486	**0.2403**
R-PREC	0.1114	0.0846	0.1191	0.1040	**0.2475**

Tableau 3 : Comparaison de notre approche par rapport à une autre approche basée sur les motifs fréquents

Le Tableau 4 montre la comparaison des résultats de la MAP, le nDCG et la F-mesure obtenus par notre approche par rapport aux différentes variantes de l'approche d'expansion proposée par les auteurs dans (Yang *et al.*, 2017), où ils ont proposé une méthode d'expansion de requêtes basée sur de multiples sources d'informations externes. Les auteurs ont utilisé le modèle *Word2Vec* que nous avons utilisé dans cet article. Nos résultats surpassent les résultats obtenus par leur approche. Nous avons obtenu des améliorations sur les trois mesures : +158,39%, +64,56% et +161,98% res-

pectivement sur la MAP, le nDCG et la F-mesure. La variante *NMF+Query* désigne la configuration utilisant une factorisation de matrice négative (NMF). La configuration *NMF+W2V* combine NMF et le modèle *Word2Vec* pour étendre la requête initiale.

Metric/RunID	NMF+Query	Word2Vec	NMF+W2V	Run-P-WE
MAP	0.036	0.093	0.027	**0.2403**
NDCG	0.226	0.219	0.272	**0.4476**
F-measure	0.101	0.039	0.092	**0.2646**

Tableau 4 : Comparaison de notre approche par rapport à une autre approche utilisant les plongements de mots

Dans notre étude expérimentale que nous avons menée, nous avons fixé le nombre de documents renvoyés, en réponse aux requêtes originales, à 500 ($N = 500$). Lorsque ce nombre augmente (jusqu'à 1000), il n'y a pas d'amélioration sur les résultats expérimentaux. Nous avons également varié la valeur du support minimum (*minsup*) pour la découverte de motifs et avons fixé ce nombre à 10 (i.e. 2%). Nous avons ainsi paramétré $k = 3$ représentant le nombre de termes les plus similaires retrouvés par la mesure de similarité cosinus. Toutes les évaluations expérimentales sont réalisées à l'aide de TREC EVAL.

6. Conclusion

Dans cet article, nous avons proposé une méthode d'expansion de requêtes permettant d'améliorer la RI dans les microblogs. La concision des microblogs ainsi des requêtes a un impact sur la qualité de recherche. Notre méthode s'appuie sur l'extraction de motifs fermés fréquents et l'exploitation de l'analyse formelle de concepts (AFC).Les motifs fermés fréquents sont combinés avec les plongements de mots pour trouver les mots les plus liés sémantiquement à la requête initiale. Cette combinaison des deux approches permet de capturer la sémantique commune d'un ensemble de tweets. Les résultats expérimentaux montrent que l'utilisation des plongements de mots avec les motifs reste meilleure que la simple utilisation de motifs. La précision a été améliorée pour la plupart des requêtes ayant des faibles résultats. Les résultats de comparaison par rapport à quelques approches de la littérature montrent que la méthode proposée améliore considérablement la précision, le rappel et la moyenne des précisions moyennes (MAP).

Nous prévoyons, en perspectives, l'exploitation des urls de tweets pour améliorer l'efficacité de la méthode proposée pour l'expansion de requêtes. Les urls peuvent avoir un contenu important qui améliore les résultats de recherche. Nous pouvons également intégrer l'information de localisation et l'aspect temporal dans la recherche des tweets où la requête peut être composée d'une région d'intérêt (ROI) et de texte. Par ailleurs, nous nous intéressons à l'utilisation des ressources externes afin de comparer la méthode proposée aux approches d'expansion externes. Enfin, d'autres expérimentations sur d'autres requêtes et collections sont également envisagées.

7. Bibliographie

Aggarwal N., Buitelaar P., « Query Expansion Using Wikipedia and Dbpedia. », *in* P. Forner, J. Karlgren, C. Womser-Hacker (eds), *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

Agrawal R., Imieliński T., Swami A., « Mining Association Rules Between Sets of Items in Large Databases », *SIGMOD Rec.*, vol. 22, n° 2, p. 207-216, June, 1993.

Anagnostopoulos I., Kolias V., Mylonas P., « Socio-semantic Query Expansion Using Twitter Hashtags », *Seventh International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2012, Luxembourg City, Luxembourg, December 3-4, 2012*, p. 29-34, 2012.

Bai J., Song D., Bruza P., Nie J.-y., Cao G., « Query expansion using term relationships language models for information retrieval », *International Conference on Information and Knowledge Management, Proceedings*, 01, 2005.

Carpineto C., Romano G., « A Survey of Automatic Query Expansion in Information Retrieval », *ACM Comput. Surv.*, vol. 44, n° 1, p. 1 :1-1 :50, January, 2012.

Codocedo V., Baixeries J., Kaytoue M., Napoli A., « Contributions to the Formalization of Order-like Dependencies using FCA », *What can FCA do for Artificial Intelligence ?*, The Hague, Netherlands, August, 2016.

Diaz F., Mitra B., Craswell N., « Query Expansion with Locally-Trained Word Embeddings », *CoRR*, 2016.

Ganter B., Wille R., *Formal concept analysis : mathematical foundations*, Springer Science & Business Media, 2012.

Hu J., Deng W., Guo J., « Improving Retrieval Performance by Global Analysis », *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 2, p. 703-706, 2006.

Jones K. S., Walker S., Robertson S. E., « A Probabilistic Model of Information Retrieval : Development and Comparative Experiments », *Inf. Process. Manage.*, vol. 36, n° 6, p. 779-808, November, 2000.

Kotov A., Zhai C., « Tapping into Knowledge Base for Concept Feedback : Leveraging Conceptnet to Improve Search Results for Difficult Queries », *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, ACM, New York, NY, USA, p. 403-412, 2012.

Kuzi S., Shtok A., Kurland O., « Query Expansion Using Word Embeddings », *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, ACM, New York, NY, USA, p. 1929-1932, 2016.

Lau C., Li Y., Tjondronegoro D., « Microblog retrieval using topical features and query expansion », *Proceedings of The Twentieth Text REtrieval Conference*, November 15-18, 2011.

Li W., Jones G. J. F., « Comparative Evaluation of Query Expansion Methods for Enhanced Search on Microblog Data : DCU ADAPT @ SMERP 2017 Workshop Data Challenge », *Proceedings of the First International Workshop on Exploitation of Social Media for Emergency Relief and Preparedness co-located with European Conference on Information Retrieval, SMERP@ECIR 2017, Aberdeen, UK, April 9, 2017.*, p. 61-72, 2017.

Massoudi K., Tsagkias M., de Rijke M., Weerkamp W., « Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts », *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, Springer-Verlag, Berlin, Heidelberg, p. 362-367, 2011.

Mikolov T., Chen K., Corrado G., Dean J., « Efficient Estimation of Word Representations in Vector Space », *Proceedings of the International Conference on Learning Representations*, p. 1-12, 2013.

Mittal N., Nayak R., Govil M. C., Jain K. C., « Dynamic Query Expansion for Efficient Information Retrieval », *2010 International Conference on Web Information Systems and Mining*, vol. 1, p. 211-215, Oct, 2010.

Nalisnick E., Mitra B., Craswell N., Caruana R., « Improving Document Ranking with Dual Word Embeddings », *Proceedings of the 25th International Conference Companion on World Wide Web*, Republic and Canton of Geneva, Switzerland, p. 83-84, 2016.

Ounis I., Amati G., Plachouras V., He B., Macdonald C., Johnson D., « Terrier Information Retrieval Platform », *Proceedings of the 27th European Conference on Advances in Information Retrieval Research*, ECIR'05, Springer-Verlag, Berlin, Heidelberg, p. 517-519, 2005.

Ounis I., Macdonald C., Lin J., Soboroff I., « Overview of the trec-2011 microblog track », *In Proceedings of TREC 2011*, 2011.

Pal D., Mitra M., Bhattacharya S., « Exploring Query Categorisation for Query Expansion : A Study », *CoRR*, 2015.

Roy D., Paul D., Mitra M., Garain U., « Using Word Embeddings for Automatic Query Expansion », *CoRR*, 2016.

Spink A., Wolfram D., Jansen J., Saracevic T., « Searching the Web : The Public and Their Queries », *Journal of the American Society for Information Science and Technology*, vol. 52, p. 226 - 234, 02, 2001.

Vulić I., Moens M.-F., « Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings », *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, ACM, New York, NY, USA, p. 363-372, 2015.

Yang Z., Li C., Fan K., Huang J., « Exploiting Multi-Sources Query Expansion in Microblogging Filtering », *Neural Network World*, vol. 27, p. 59-76, 01, 2017.

Zaki M. J., Hsiao C., « Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure », *IEEE Trans. Knowl. Data Eng.*, vol. 17, n° 4, p. 462-478, 2005.

Zhai C., Lafferty J., « Model-based Feedback in the Language Modeling Approach to Information Retrieval », *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, ACM, New York, NY, USA, p. 403-410, 2001.

Zheng G., Callan J., « Learning to Reweight Terms with Distributed Representations », *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, ACM, New York, NY, USA, p. 575-584, 2015.

Zingla M. A., Chiraz L., Slimani Y., « Short Query Expansion for Microblog Retrieval », *Knowledge-Based and Intelligent Information & Engineering Systems : Proceedings of the 20th International Conference KES 2016*, vol. 96, p. 225-234, October, 2016.

Zingla M. A., Latiri C., Mulhem P., Berrut C., Slimani Y., « Hybrid Query Expansion Model for Text and Microblog Information Retrieval », *Inf. Retr.*, vol. 21, n° 4, p. 337-367, August, 2018.