Intégration des facteurs temps et autorité sociale dans un modèle bayésien de recherche de tweets

Lamjed Ben Jabeur, Lynda Tamine et Mohand Boughanem

IRIT, Université Paul Sabatier
118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9
{jabeur, tamine, boughanem}@irit.fr

RÉSUMÉ. *Cet article présente une approche sociale pour la recherche d'information dans les microblogs intégrant diverses sources d'évidence au sein d'un réseau bayésien. Notre contribution consiste à étendre la notion classique de pertinence, basée sur la similarité textuelle, par de nouveaux facteurs tels que l'importance sociale des blogueurs et la magnitude temporelle des microblogs. Dans ce papier, l'importance sociale d'un blogueur est assimilée à son influence dans le réseau et est évaluée par un score de PageRank déduit sur le réseau de diffusion des microblogs. Nous proposons d'estimer la magnitude temporelle selon le nombre de voisins temporels qui incluent les termes de la requête. Afin de valider notre approche, une évaluation expérimentale à été menée sur la collection de microblogs de référence TREC Tweets2011. Les résultats montrent que notre modèle présente un gain de* 24% *par rapport à la médiane des résultats officiels de TREC Microblog 2011.*

ABSTRACT. *We present in this paper a social-based approach for information retrieval in microblogs that integrates various sources of evidences using Bayesian networks. In addition to textual similarly, we propose to integrate new relevance factors such as the social importance of microbloggers and the time magnitude of microblogs. In this work, the social importance of a microblogger is assimilated to his influence on the social network and computed by applying PageRank algorithm on the retweet network. We propose to estimate the time magnitude from the temporal neighbors that include query terms. To validate our approache, we conduct a series of experiments on the TREC Tweets2011 dataset. Results show that our model overpass the median of official runs with an improvement of* 24%.

MOTS-CLÉS : *Microblogs, recherche de tweets, réseaux sociaux, influence, magnitude temporelle*

KEYWORDS: *Microblogs, tweet search, social network, influence, time magnitude*

CORIA 2012, pp. 301–316, Bordeaux, 21-23 mars 2012

1 Introduction

Un service de *microblogage* est à la fois un moyen de communication et un système de collaboration qui permet le partage et la diffusion de messages textuels. En comparaison avec les autres réseaux sociaux sur le *Web*, les articles de *microblog* sont particulièrement courts et soumis en temps réel pour rapporter un événement récent. Dans ce travail, nous nous intéressons au service de *microblogage* de *Twitter*[1], étant le service le plus populaire et le plus largement utilisé. *Twitter* se distingue de sites similaires par certaines caractéristiques et fonctionnalités. Une des caractéristiques importantes est la présence de la relation sociale d'abonnement. Cette relation directionnelle permet aux utilisateurs d'exprimer leur intérêt pour les publications d'un blogueur particulier. De plus, *Twitter* se distingue par le principe de diffusion des articles, qui sont appelés dans le contexte de ce service *"tweets"*. Grâce à cette fonctionnalité, l'utilisateur peut transmettre un *tweet* intéressant à ses abonnés. Le message rediffusé est appelé *"retweet"*. Un blogueur, appelé aussi *twitterer*, peut annoter ses *tweets* en utilisant des *#hashtags* ou l'adresser à un utilisateur spécifique à travers les *mentions @utilisateur*. Enfin, un *tweet* permet aussi de partager une ressource Web référencée par une URL. La figure 1 résume les principales entités impliquées dans le réseau social de *Twitter* et les diverses relations qui les associent.

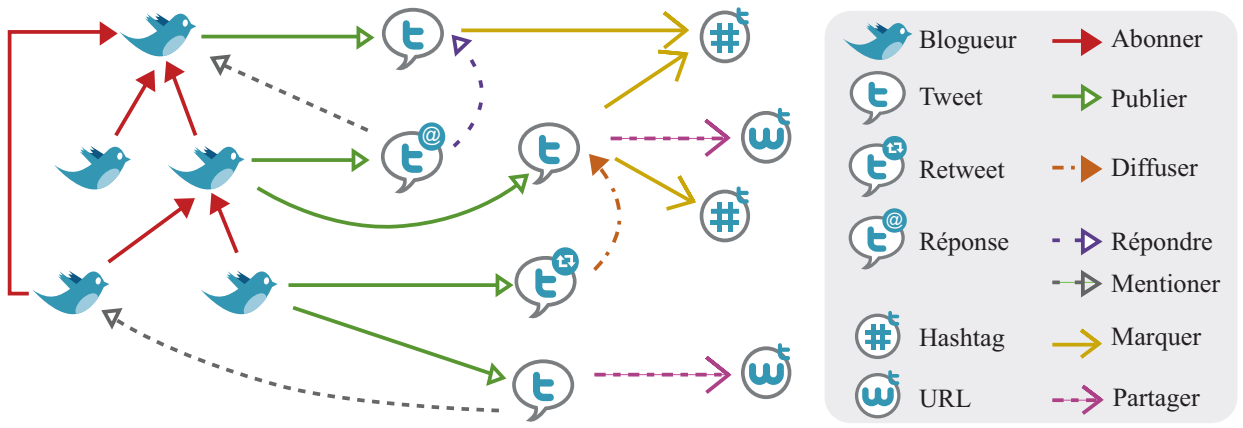


Figure 1 – Réseau social d'information de Twitter

Selon les statistiques officielles publiées en mars 2011[2], environ 140 millions de *tweets* sont soumis chaque jour sur *Twitter*. Avec ce taux important, les données générées par les *microblogs* sont de plus en plus disponibles. Cependant, les utilisateurs sont submergés par l'énorme quantité de nouveaux *tweets* et trouvent une difficulté à accéder aux publications qui les intéressent. Une nouvelle tâche de *recherche de tweets* est donc nécessaire. Cette tâche est définie par une fonction $RSV(q, t_i, \theta_q)$ qui évalue la pertinence d'un *tweet* t_j par rapport à une requête q tout en tenant compte de la date de soumission de la requête θ_q. La campagne d'évaluation *TREC 2011* définit la recherche de *tweets* comme une tâche de recherche en temps réel où l'utilisateur souhaite accéder aux publications les plus pertinentes à une requête, mais aussi les publications les plus récentes (Ounis *et al.,* 2011) . Autrement dit, cette tâche consiste

1. http ://www.twitter.com/
2. http ://blog.twitter.com/2011/03/numbers.html

à trouver les *tweets* sur un sujet X à un instant t. A la différence de la recherche Web traditionnelle, la recherche de *tweets* permet de sélectionner une information brève, concise et récente sur un sujet ou un évènement survenu récemment. Vu la spécificité de *microblogs*, la recherche de *tweets* est confrontée à plusieurs défis tels que l'indexation du flux des articles (Sankaranarayanan *et al.,* 2009), la détection des *spams* (Yardi *et al.,* 2010), la diversification des résultats (Choudhury *et al.,* 2011) et l'évaluation de la qualité des *tweets* (Pal *et al.,* 2011, Nagmoti *et al.,* 2010).

Dans ce papier, nous proposons un modèle pour la recherche de *tweets* qui évalue la qualité des *tweets* selon deux contextes, social et temporel. Nous proposons d'estimer la qualité d'un *tweet* par l'autorité sociale du blogueur correspondant. Cette autorité est calculée par l'application de l'algorithme *PageRank* sur le réseau d'influence sociale. Dans le même objectif, la qualité d'un *tweet* est évaluée selon sa date de publication. Les *tweets* soumis dans les périodes d'activité d'un terme de la requête sont alors caractérisés par une plus grande importance. Enfin, nous proposons d'intégrer l'importance sociale du blogueur et la magnitude temporelle avec les autres facteurs de pertinence en utilisant un modèle bayésien de croyance.

Cet article est organisé comme suit. La section 2 présente un aperçu des travaux proposés dans les domaines de la recherche de *tweets* et de l'évaluation de l'importance sociale des blogueurs. Nous introduisons dans la section 3 notre modèle bayésien pour la recherche de *tweets*. Nous décrivons dans la section 4 le processus d'évaluation de la requête et le calcul des probabilités conditionnelles. Dans la section 5, nous menons une série d'expérimentations sur la collection de référence TREC Microblog 2011 pour valider notre approche. Enfin, la section 6 conclut l'article et annonce des perspectives.

2 Recherche de tweets pertinents : synthèse des travaux

La recherche d'information dans les *microblogs* a été introduite pour la première fois par les travaux de (Java *et al.,* 2007). Ces travaux se focalisent sur l'analyse des données et de la structure du réseau social. Plus récemment, certains travaux ont abordé la recherche de *tweets* comme un problème indépendant qui se distingue de la recherche Web classique (Teevan *et al.,* 2011). Ces approches proposent d'utiliser diverses sources d'évidence telles que le contexte spatio-temporel, les caractéristiques des *microblogs* et la structure du réseau social.

Les approches basées sur le contexte spatio-temporel proposent de regrouper les *tweets* selon des classes géographiques puis les agréger pour construire un flux d'actualités locales (Sankaranarayanan *et al.,* 2009). Dans la même catégorie d'approches spatio-temporelles, certains travaux proposent d'identifier les périodes d'activité d'un sujet donné, puis de sélectionner les *tweets* les plus représentatifs de chaque période (Grinev *et al.,* 2009). La deuxième catégorie d'approches propose d'utiliser les caractéristiques des *microblogs* comme une alternative à l'ordonnancement chronologique des *tweets*. Ces caractéristiques regroupent certaines fonctionnalités et données telles que le nombre d'abonnements, de *tweets*, de *retweets*, de réponses, de *hashtags* et d'URLs (Nagmoti *et al.,* 2010). Enfin, certaines approches sociales proposent d'éva-

luer la qualité d'un *tweet* selon la position du blogueur correspondant dans le réseau social. Les résultats sont ainsi classés en tenant compte de la structure du réseau et en évaluant l'autorité du blogueur (Duan *et al.,* 2010).

Les approches sociales pour la recherche de *tweets* évaluent l'importance des blogueurs selon différentes mesures. Dans la première catégorie de mesures, les indicateurs d'importance sont estimés à partir des données quantitatives. Parmi les mesures proposées, nous citons le nombre d'abonnés (*Indegree influence*), le nombre des *retweets* (*Retweet influence*), le nombre de mentions (*Mention influence*) (Cha *et al.,* 2010), le nombre de *tweets* (*TweetRank*) et le nombre proportionnel d'abonnés (*FollowerRank*) (Nagmoti *et al.,* 2010). Par ailleurs, les approches de classification proposent de classer les blogueurs dans des clusters thématiques. Ensuite, les éléments de chaque cluster sont triés par autorité afin de sélectionner les blogueurs les plus importants dans le réseau (Pal *et al.,* 2011). Enfin, certaines approches sociales exploitent les liens du réseau afin de mesurer l'importance des blogueurs. Parmi les mesures proposées, nous citons *le score d'influence* calculé par l'application de l'algorithme *PageRank* sur le réseau social d'abonnement (Kwak *et al.,* 2010). Cette mesure a été étendue par l'algorithme *TwitterRank* (Weng *et al.,* 2010) qui calcule un *PageRank* thématique sur le réseau d'abonnement. De même, *le score de popularité* (Duan *et al.,* 2010) mesure l'importance du blogueur par l'application de l'algorithme *PageRank* sur le réseau social de *retweets*.

Nous proposons dans ce papier une approche sociale pour la recherche de *tweets*, qui tient compte de l'importance des blogueurs et de la magnitude temporelle des *tweets*. Comparativement à nos travaux précédents basés sur les réseaux d'inférence Bayésiens (Jabeur *et al.,* 2012), nous présentons ici une *nouvelle topologie du modèle bayésien basée sur réseau de croyance*. *De plus,* nous étudions l'impact des facteurs temps et autorité sociale sur l'efficacité de recherche et nous menons une nouvelle série d'expérimentations avec la collection de référence TREC Microblog 2011. Comparativement aux approches dans le domaine, nous proposons :

– d'intégrer les différentes sources d'évidence en utilisant un modèle de réseau bayésien. Les travaux similaires proposent de les combiner linéairement (Nagmoti *et al.,* 2010) ou d'apprendre une fonction d'ordonnancement avec de multiples facteurs en entrée (Duan *et al.,* 2010) ;

– de modéliser le réseau social des blogueurs avec les relations des *retweets* et non du réseau d'abonnement largement utilisé dans littérature (Kwak *et al.,* 2010, Weng *et al.,* 2010). Dans ce cas, l'importance du blogueur est assimilée à son influence dans le réseau. Nous considérons uniquement le sous-réseau des blogueurs généré par la liste des résultats afin d'éviter la domination de certaines célébrités dans le réseau global des *retweets* (Duan *et al.,* 2010) ;

– d'estimer la magnitude temporelle à partir de l'occurrence de chaque terme séparé dans le voisinage temporel à la différence d'autres travaux qui étudient la distribution des *tweets* dans le temps (Grinev *et al.,* 2009).

3 Un modèle bayésien pour la recherche de tweets

La recherche de *tweets* est une tâche particulière, motivée par une variété de besoins thématiques, sociaux et temporels (Teevan *et al.,* 2011). Afin d'y répondre, il est nécessaire d'intégrer plusieurs sources d'évidence qui peuvent être dépendantes entre elles. Tenant compte de cette spécificité, nous proposons de modéliser la recherche de *tweets* par les réseaux bayésiens. Cette famille de modèles permet en effet de représenter la dépendance entre les sources d'évidence par des probabilités conditionnelles. De plus, les réseaux bayésiens permettent d'estimer les données manquantes et d'assurer la tâche de recherche même si certaines données ne sont pas disponibles telles que les profils protégés des blogueurs. Dans cette section, nous introduisons quelques définitions et notations puis nous décrivons la topologie du réseau bayésien proposé.

3.1 Définitions et notations

- Réseau bayésien : Un réseau bayésien représente un formalisme graphique qui permet de représenter des relations causales entres des variables aléatoires. Un réseau bayésien est un graphe dirigé acyclique $G(X, E)$, avec l'ensemble des nœuds X correspondant à des variables aléatoires et l'ensemble des arcs $E = X \times X$ représentant les dépendances conditionnelles entre elles. Soit X_i une variable aléatoire et $Pa(X_i)$ l'ensemble de ses parents. La probabilité jointe des toutes les variables du réseau est définie par $P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i))$.

- Termes : L'ensemble des termes présents dans l'index constitue l'univers de discours $U = k_1, k_2, \ldots, k_n$. Chaque terme k_i est associé à une variable aléatoire $k_i \in \{0, 1\}$. La représentation abrégée de $k_i = 1$ est notée k_i et signifie que *"le terme k_i est observé"*. De la même manière, $k_i = 0$ est noté $\bar{k_i}$ et signifie que *"le terme k_i n'est pas observé"*. Cette même notation (k_i) est utilisée pour désigner à la fois le terme, la variable aléatoire et le nœud correspondant dans le réseau. Pour n termes dans l'index, il existe 2^n configurations possibles qui permettent de représenter un *tweet* ou une requête utilisateur. Dans le cas d'un index contenant deux termes, l'ensemble des configurations possibles est $\{(k_1, k_2), (k_1, \bar{k_2}), (\bar{k_1}, k_2), (\bar{k_1}, \bar{k_2})\}$. Soit une configuration $\vec{k}$. La fonction $on(k_i, \vec{k})$ associe chaque variable k_i à sa valeur correspondante dans k. $on(k_i, k) = 1$ si et seulement si le terme k_i est instancié dans $\vec{k}$.

- Tweets : Les *tweets* remplacent les documents dans le modèle bayésien de base. Chaque tweet t_j est associé à une variable aléatoire $t_j \in \{0, 1\}$. La représentation abrégée de $t_j = 1$ est notée t_j. $t_j = 1$ signifie que *"le tweets t_j est observé"* et inversement $t_j = 0$ est noté $\bar{t_j}$. Chaque *tweet* t_j est représenté par un ensemble de termes $t_j = k_1, ..., k_i, ..., k_n$ avec k_i est une variable aléatoire qui indique la présence du terme k_i dans le *tweet* t_j. En outre, nous proposons d'associer à chaque *tweet* t_j trois variables aléatoires t_{kj}, t_{oj} et t_{sj} selon trois contextes différents. La première variable t_{kj} correspond à la probabilité d'observer le *tweet* selon l'évidence thématique. La seconde variable t_{oj} correspond à la probabilité d'observer le *tweet* avec la

connaissance implicite du contexte temporel. Enfin, t_{sj} correspond à la probabilité d'observer le *tweet* avec la connaissance implicite du contexte social. Ces probabilités permettent de décomposer l'événement et d'observer un document selon les trois évidences : thématique, temporelle et sociale.

- Périodes : Une période o_e correspond à une fenêtre temporelle de taille Δt exprimée en une unité temporelle λt. Chaque période couvre un intervalle temporel $\left[\theta_{o_e} - \frac{\Delta t}{2}, \theta_{o_e} + \frac{\Delta t}{2}\right]$ avec θ_{o_e} étant la date de la période o_e qui correspond au centre de sa fenêtre temporelle. Une variable aléatoire $o_e \in \{0, 1\}$ est associée à chaque période. $o_e = 1$, noté aussi o_e , signifie que *"la période o_e est sélectionnée"*. Inversement, $o_e = 0$, noté $\bar{o}_e$, signifie que *"la période o_e n'est pas sélectionnée"*.

- Blogueurs : Chaque blogueur est représenté par un nœud u_f dans le réseau. Une variable aléatoire $u_f \in \{0, 1\}$ est associée à chaque nœud blogueur. $u_f = 1$, noté u_f, signifie que *"le blogueur u_f est observé"*. Inversement, $u_f = 0$, noté $\bar{u}_f$, signifie que *"le blogueur u_f n'est pas observé"*.

3.2 Topologie du réseau de croyance

Dans cette section, nous décrivons la topologie de notre réseau bayésien de recherche de *tweets*. Cette topologie repose sur les travaux de (Silva *et al.,* 2000) proposant de combiner au sein d'un réseau bayésien de croyance les évidences liées au contenu et celles liées aux liens hypertextes. Contrairement aux réseaux d'inférence où l'unique racine du réseau est représentée par la requête, les racines du réseau bayésien de croyance correspondent aux nœuds termes.

Un terme k_i appartenant à l'index de *tweets* est représenté par un nœud k_i dans le réseau. L'ensemble de ces nœuds constitue la couche des termes U. Une requête utilisateur est modélisée par un noeud q associé à une variable aléatoire $q \in \{0, 1\}$. Il existe un arc (k_i, q) entre le nœud requête et chaque terme parent k_i tel que $on(k_i, q) = 1$. Chaque *tweet* t_j est représenté par trois nœuds t_{kj}, t_{oj} et t_{sj} appartenant respectivement à la couche d'évidence thématique TK, la couche d'évidence temporelle TO et la couche d'évidence sociale TS. Ces trois nœuds sont liés à un quatrième nœud t_j qui représente également le *tweet* t_j. Il existe alors un arc (t_{xj}, t_j) entre t_{xj} et t_j quelque soit $x \in \{k, o, s\}$. L'ensemble des nœuds t_j constitue la couche de *tweets* T. Parmi ces nœuds, seulement t_{kj} est lié aux nœuds termes. Ainsi, un arc (k_i, t_{kj}) est défini entre t_{kj} et chaque terme k_i présent dans le *tweet* $on(k_i, t_j) = 1$.

Chaque période o_e est représentée par un nœud o_e dans le réseau. L'ensemble de ces nœuds constitue la couche des périodes O. Les nœuds périodes sont liés à la couche d'évidence temporelle TO. Un arc (o_e, t_{oj}) est défini entre o_e et t_{oj} si et seulement si la date de publication de *tweet* t_j est incluse dans sa fenêtre temporelle $\left|\theta_{t_j} - \theta_{o_e}\right| \leq \frac{\Delta t}{2}$ et si le centre de cette période θ_{o_e} est le plus proche à θ_{t_j} : $|\theta_{t_j} - \theta_{o_e}| = min_{\forall o_i}|\theta_{t_j} - \theta_{o_i}|$. Une période o_e est également liée par un arc (k_i, o_e) à un nœud terme k_i si et seulement si le terme apparait dans sa fenêtre temporelle $\{k_i | \forall t_j, on(k_i, t_j) = 1 \wedge \left|\theta_{t_j} - \theta_{o_e}\right| \leq \frac{\Delta t}{2}\}$.

Chaque blogueur u_f est représenté par un nœud dans le réseau. Ce nœud est lié aux *tweets* correspondants dans la couche d'évidence sociale TS. Il existe alors un arc (u_f, t_{sj}) entre le blogueur u_f et chaque nœud t_{sj} si le *tweet* t_j est publié par u_f.

La figure 2 illustre la topologie du réseau de croyance pour la recherche de *tweets*.

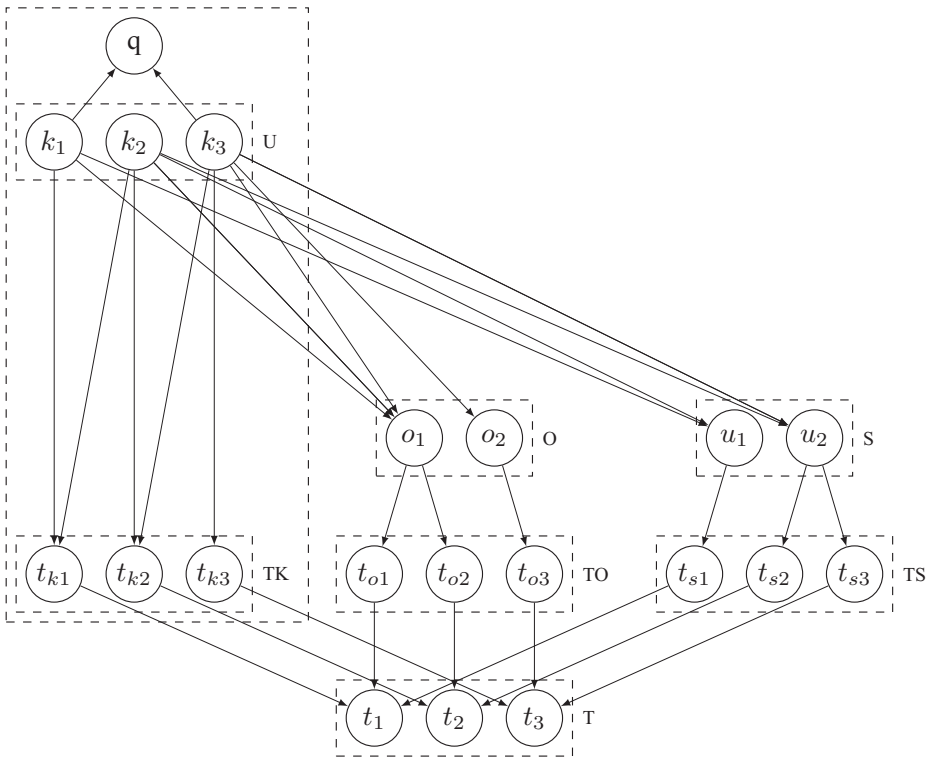


Figure 2 – Réseau de croyance bayésien pour la recherche de *tweets*

4 Évaluation de la requête

4.1 Principe général

La pertinence d'un *tweet* t_j étant donnée une requête q soumise à une date θ_q est exprimée par la probabilité $P(t_j|q, \theta_q)$. Indépendamment de la date de requête, cette probabilité peut être estimé par $P(t_j|q) = \frac{P(t_j \wedge q)}{P(q)}$. La probabilité $P(q)$ a une valeur constante pour tous les *tweets* de la collection. Ainsi, $P(t_j|q)$ est approximé à $P(t_j|q) \propto P(t_j \wedge q)$. En se basant sur la topologie du réseau bayésien décrite dans la figure 2, la probabilité $P(t_j|q)$ est estimée par la formule suivante :

$$P(t_j|q) \propto \sum_{\forall \vec{k}} P(q|\vec{k}) P(t_j|\vec{k}) P(\vec{k}) \quad [1]$$

$\vec{k}$ est une configuration des termes inclus dans l'univers de discours U.

La probabilité $P(t_j|\vec{k})$ dépend des trois composantes : l'évidence thématique, l'évidence temporelle et l'évidence sociale. Cette probabilité est estimée par combinaison de trois probabilités avec l'opérateur de conjonction *AND*. Ainsi :

$$P(t_j|\vec{k}) = P(t_{kj}|\vec{k}) P(t_{oj}|\vec{k}) P(t_{sj}|\vec{k}) \quad [2]$$

En substituant $P(t_j|\vec{k})$ dans la formule 1, la pertinence d'un *tweet* peut être approximée par la formule suivante :

$$P(t_j|q) \propto \sum_{\forall \vec{k}} P(q|\vec{k})P(t_{kj}|\vec{k})P(t_{oj}|\vec{k})P(t_{sj}|\vec{k})P(\vec{k}) \quad [3]$$

Afin de respecter les exigences temporelles de la tâche de recherche de *tweets*, nous proposons d'éliminer tout *tweet* t_j dont la date de publication θ_{t_j} est postérieure à la date de soumission de la requête θ_q. Nous donnons une valeur de probabilité nulle $P(t_j|q) = 0$ pour chaque *tweet* t_j dont $\theta_{t_j} > \theta_q$.

4.2 Estimation des probabilités conditionnelles

Dans cette section, nous détaillons le principe de calcul des probabilités introduites dans la formule 3 et nous discutons les différentes possibilités de les estimer.

4.2.1 Probabilité $P(\vec{k})$

La probabilité $P(\vec{k})$ correspond à la probabilité d'apparition de la configuration $P(\vec{k})$ dans l'univers U. Nous supposons ici l'indépendance des termes et l'équivalence des probabilités d'apparition pour toutes configurations. Ayant n termes dans la collection, cette probabilité est estimée comme suit :

$$P(\vec{k}) = \frac{1}{2^n} \quad [4]$$

4.2.2 Probabilité $P(q|\vec{k})$

La probabilité que la requête q couvre la configuration $\vec{k}$, notée $P(q|\vec{k})$, pondère les différentes configurations possibles dans l'univers U. Nous proposons d'estimer cette probabilité comme suit :

$$P(q|\vec{k}) = \begin{cases} 1, & \text{si } on(k_i, q) = on(k_i, \vec{k}), \forall k_i \\ 0, & \text{sinon} \end{cases} \quad [5]$$

Notons que $P(q|\vec{k}) > 0$ si et seulement si tous les termes de la requête sont instanciés dans la configuration $\vec{k}$ et inversement pour le reste des termes. $P(\bar{q}|\vec{k}) = 1 - P(q|\vec{k})$.

4.2.3 Probabilité $P(t_j|\vec{k})$

La probabilité que le *tweet* t_j couvre la configuration $\vec{k}$, notée $P(t_j|\vec{k})$, mesure la similarité entre le *tweet* et la configuration $\vec{k}$. D'une manière générale, cette probabilité est estimée selon la fréquence des termes dans le *tweet* tf_{k_i,t_j}. Cependant, les termes sont rarement répétés dans le même *tweet* vu la limitation sur le nombre des

caractères. En tenant compte de cette spécificité des *tweets* et en supposant que tous les termes ont la même importance, la probabilité $P(t_j|\vec{k})$ est estimée par :

$$P(t_{kj}|\vec{k}) = \frac{1}{i(\vec{k})} \sum_{\substack{\forall k_i \in \vec{k}, on(k_i,\vec{k})= \\ on(k_i,t_j)=1}} \frac{tf_{k_i,t_j} - \beta}{tf_{k_i,t_j}} \quad [6]$$

Le quotient $\frac{tf_{k_i,t_j}-\beta}{tf_{k_i,t_j}}$ permet de transformer les grandes valeurs de fréquence en un intervalle plus réduit. $\beta \in [0,1]$ est une paramètre de transformation. Plus la valeur de β est petite, moins d'importance est donnée aux fréquences élevées et inversement. $i(\vec{k}) = |\{k_i \in k : on(k_i,k) = 1\}|$ correspond au nombre de termes instanciés positivement dans la configuration $\vec{k}$. La probabilité complémentaire est estimée par $P(\bar{t}_{kj}|\vec{k}) = 1 - P(t_{kj}|\vec{k})$.

4.2.4 Probabilité $P(t_{oj}|\vec{k})$

La probabilité $P(t_{oj}|\vec{k})$ exprimée en fonction de o_e parent du nœud t_{oj} est :

$$P(t_{oj}|\vec{k}) = P(t_{oj}|o_e)P(o_e|\vec{k}) + P(t_{oj}|\bar{o}_e)P(\bar{o}_e|\vec{k}) \quad [7]$$

Or la probabilité d'observer un *tweet*, sachant que sa période correspondante n'est pas sélectionnée, est nulle $P(t_{oj}|\bar{o}_e) = 0$. Alors, la probabilité $P(t_{oj}|\vec{k})$ devient :

$$P(t_{oj}|\vec{k}) = P(t_{oj}|o_e)P(o_e|\vec{k}) \quad [8]$$

La probabilité $P(t_{oj}|o_e)$ d'observer un *tweet* t_j étant donnée une période o_e permet de pondérer les différents *tweets* d'une période donnée. Notons que la visibilité d'un *tweet* augmente avec le nombre des *retweets* reçus. Ainsi, nous proposons de calculer cette probabilité par la proportion des *retweets* sur le nombre total des *tweets* publiés :

$$P(t_{oj}|o_e) = \frac{1 + |\rho_{o_e}(t_j)|}{|\tau(o_e)|} \quad [9]$$

$\rho_{o_e}(t_j)$ est l'ensemble des *tweets* publiés par t_j durant la période o_e et diffusés (*retweets*) par un autre blogueur durant la même période. $\tau(o_e)$ est l'ensemble des *tweets* publiés durant la période o_e.

La probabilité $P(o_e|\vec{k})$ d'arriver sur une période o_e étant donnée une configuration $\vec{k}$ permet d'attribuer des poids à chaque période. Nous proposons d'estimer cette probabilité selon la fréquence des *tweets* dans la période o_e donnée par la formule suivante :

$$P(o_e|\vec{k}) = \frac{\sum_{\forall k_i, on(k_i,\vec{k})=1} df_{k_i,o_e}}{\sum_{\forall k_i, on(k_i,\vec{k})=1} df_{k_i}} \quad [10]$$

df_{k_i,o_e} est le nombre de *tweets* contenant le terme k_i et publiés dans la période o_e. df_{k_i} est le nombre total de *tweets* contenant le terme k_i.

4.2.5 Probabilité $P(t_{sj}|\vec{k})$

La probabilité $P(t_{sj}|\vec{k})$ peut être développée en fonction de u_f comme suit :

$$P(t_{sj}|\vec{k}) = P(t_{sj}|u_f)P(u_f|\vec{k}) + P(t_{sj}|\bar{u}_f)P(\bar{u}_f|\vec{k}) \quad [11]$$

Or la probabilité d'observer un *tweet*, sachant que le blogueur correspondant n'est pas observé, est nulle $P(t_{sj}|\bar{u}_f) = 0$. Alors, la probabilité $P(t_{sj}|\vec{k})$ devient : $P(t_{sj}|\vec{k}) = P(t_{sj}|u_f)P(u_f|\vec{k})$. Supposons que les deux variables aléatoires u_f et $\vec{k}$ sont indépendantes, cette probabilité est alors estimée comme nuit :

$$P(t_{sj}|\vec{k}) = P(t_{sj}|u_f)P(u_f) \quad [12]$$

La probabilité $P(t_{sj}|u_f)$ d'obtenir le *tweet* t_j étant donnée le blogueur u_f permet de pondérer les différents *tweets* de chaque blogueur. Nous proposons d'estimer cette probabilité d'une manière uniforme pour l'ensemble de *tweets* $\tau(u_f)$ publiés par u_f :

$$P(t_{sj}|u_f) = \frac{1}{|\tau(u_f)|} \quad [13]$$

La probabilité $P(u_f)$ d'observer un blogueur u_f reflète son importance dans le réseau social. Un blogueur reçoit une grande importance s'il influence son réseau et si les *tweets* correspondants sont largement diffusés (retweet). Dans ce travail, l'importance sociale est interprétée par l'influence du blogueur et est estimée en appliquant l'algorithme *PageRank* sur le réseau social des *retweets*. Pour éviter la domination de certaines célébrités caractérisée par un nombre élevé des *retweets*, nous proposons d'appliquer l'algorithme *PageRank* uniquement sur le sous-réseau des blogueurs généré par les *tweets* instanciés par la requête. Cela permet d'évaluer l'influence d'un blogueur pour un sujet spécifique.

Le réseau social des blogueurs est modélisé par un graphe $G = (U, R)$ où U représente l'ensemble des blogueurs instancié dans le réseau bayésien et $R = U \times U$ désigne l'ensemble des relations de *retweet*. Un blogueur u_i est inclus dans le réseau si l'un de ses *tweets* contient au moins un terme de la requête. Une relation de *retweet* $(u_i, u_j) \in R$ est définie à partir d'un blogueur u_i vers u_j si et seulement s'il existe au moins un *tweet* publié par u_j et rediffusé par u_i. Un poids d'influence est affecté à chaque association de *retweet* en fonction de l'ensemble des *tweets* rediffusés par u_j, noté $\rho(u_i)$, et l'ensemble des *tweets* $\tau(u_j)$ publiés par u_j :

$$w_{i,j} = \frac{|\tau(u_j) \wedge \rho(u_i)|}{|\rho(u_i)|} \quad [14]$$

Le score d'influence d'un *blogueur* est calculé sur plusieurs itérations comme suit :

$$Inf_G^{p+1}(u_i) = d\frac{1}{|U|} + (1-d) \sum_{u_j : e(u_j,u_i) \in E} w_{j,i} \frac{Inf_G^p(u_j)}{O(u_j)} \quad [15]$$

$O(u_j)$ est le nombre de relations de source u_j. $d \in [0, 1]$ est un paramètre de configuration de l'algorithme *PageRank*. $w_{j,i}$ est le poids de la relation d'influence entre u_j

et u_i comme défini dans l'équation 14. Enfin, les scores d'influence sont normalisés par rapport à la somme des scores de tous les blogueurs.

5 Évaluation expérimentale

Nous menons une série d'expérimentations afin d'évaluer notre modèle pour la recherche de *tweets*. Les deux principaux objectifs de cette évaluation sont : (1) étudier l'impact de chaque composante de pertinence notamment les facteurs temps et autorité sociale sur l'efficacité de recherche ; (2) comparer notre approche aux modèles traditionnels de recherche d'information et aux approches similaires pour la recherche de *tweets*, sur la base des résultats officiels TREC Microblog 2011.

5.1 Description du cadre expérimental

- Collection des tweets : Ces expérimentations exploitent le corpus *tweets2011* distribué dans cadre la campagne d'évaluation TREC Microblog 2011 (Ounis *et al.,* 2011). Cette collection inclut environ 16 millions de *tweets* collectés sur 16 jours. Le tableau 1 présente les statistiques générales du corpus. Cette collection est indexée en utilisant la plateforme de recherche de *tweets* NESTOR développée au sein de notre équipe. Ce système de recherche prend en compte les principales caractéristiques des *tweets*, telles que les mentions et les hashtags, et détecte la forme déclarative des *retweets "RT @mention"*. En outre, ce système distingue les URLs et les adresses emails et indexe seulement le contenu textuel des *tweets*. De plus, ce système applique une technique de tokenisation multilingue sur les textes et intègre un algorithme à N-grammes pour l'indentification des langues (Cavnar *et al.,* 1994). L'algorithme de Lemmatisation *Porter Stemming* est appliqué sur les textes en anglais. Enfin, la recherche de *tweets* est assurée en respectant les exigences de tâche de recherche en temps réel de TREC Microblogs 2011.

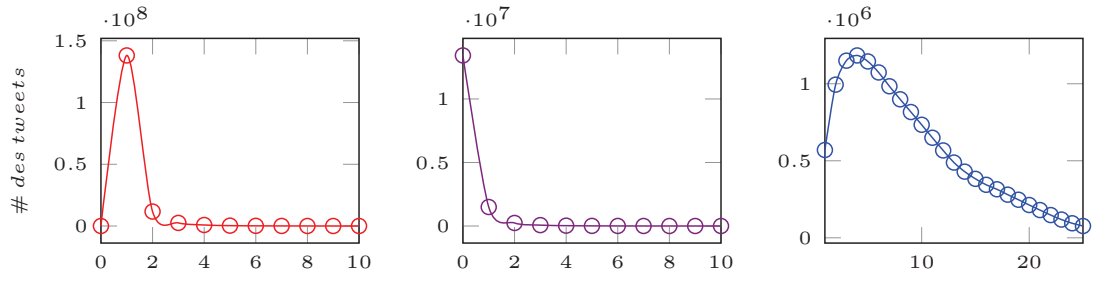
Tweets	16141812	Blogueurs	5356432
Retweets	1128179	Relations des retweets	1060551
Tweet avec hashtags	1860112	Nœuds dans le réseau social des retweets	5495081
Termes uniques	7781775	Arcs dans le réseau social des retweets	1024914
Hashtags uniques	455179	Composante la plus géante du réseau	11.12%

Tableau 1 – Statistique sur la collection

Le réseau social construit à travers les relations de *retweet* entre les blogueurs inclut environ 5.5 millions de nœuds. Le nombre des nœuds blogueurs présenté dans le tableau 1 dépasse le nombre des blogueurs dans la collection puisque certains *retweets* pointent sur d'autres utilisateurs en dehors de la collection. Enfin, la composante connexe la plus géante du réseau inclut environ $11,12\%$ des nœuds.

La figure 3 présente la distribution des fréquences des termes, des hashtags et des longueurs des *tweets*. La figure 3.a présente la distribution des *tweets* en fonction des fréquences. Cette analyse montre que les termes apparaissent souvent qu'une seule fois dans les mêmes *tweets*. La figure 3.b relative à la distribution des hashtags montre

que la plupart des *tweets* (88%) ne présentent aucun hashtag. Selon la figure 3.c, nous constatons que la distribution des *tweets* atteint une valeur maximale dans le cas des courts *tweets* à 4 termes. Cependant, un tweet inclut 11 termes en moyenne.



(a) Fréquence des termes (b) Hashtags (c) Longueur des tweets

Figure 3 – Distributions des fréquences, des hashtags et des longueurs des *tweets*

- Tâche de recherche : Selon la définition de TREC Microblog 2011, la tâche de recherche en temps réel vise à retrouver les *tweets* les plus pertinents et également les plus récents pour une requête q soumise une date donnée. La collection *tweets2011* inclut 49 requêtes avec les dates correspondantes. Les jugements de pertinence sont construits à partir des 30 premiers résultats des 184 systèmes participants. Une valeur de pertinence graduelle entre 0 et 2 est attribuée à chaque tweet. Cependant, les *retweets* sont automatiquement considérés non pertinents et seulement les *tweets* en anglais sont analysés. Afin de respecter la contrainte temporelle, aucune source d'information postérieure à la date de la requête ne peut être utilisée y compris les *tweets* dans la collection qui sont publiés après la date de la requête. Le classement final des résultats est établi selon l'ordre chronologique inverse contrairement aux autres tâches TREC qui classent les documents selon leurs scores. Vu la faible densité des documents pertinents, l'ordonnancement des *tweets* par date de publication engendre la "dispersion" des *tweets* pertinents. Il est alors nécessaire de réduire le bruit apporté par les documents récents et de présenter uniquement les *tweets* de haute qualité. Pour cela nous proposons de filtrer la liste initiale des résultats ordonnés en fonction des scores selon les conditions suivantes : (1) supprimer les *tweets* écrits en une langue autre que l'anglais ; (2) supprimer tous les *tweets* de conversation qui commencent par *"@mention"* ; (3) supprimer les *tweets* au-delà de la $30^{ème}$ position.

- Mesures d'évaluation : la précision $p@30$ est la mesure officielle pour l'évaluation de la tâche de recherche en temps réel dans TREC Microblog 2011. Cette mesure évalue la capacité d'un système à retourner les *tweets* pertinents parmi les 30 premiers de la liste des résultats. La précision moyenne *MAP* est utilisée comme une mesure supplémentaire pour évaluer l'efficacité de recherche tout en tenant compte de la précision, du rappel et du rang des documents.

- Modèles de référence : Le tableau 2 présente les différentes configurations et *baselines* utilisées comme référence dans cette évaluation. Ce tableau indique aussi la technique de filtrage utilisée, notamment le seuil appliqué sur le nombre des résultats.

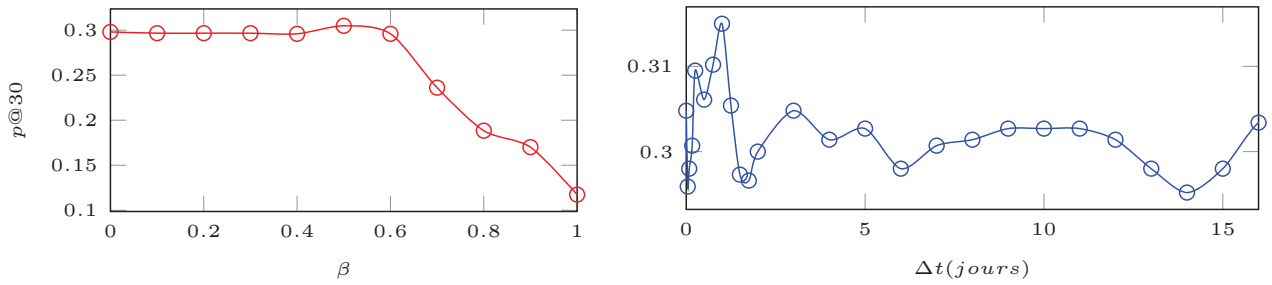
BNTS	△		Notre modèle bayésien de croyance pour le recherche des tweets
BNTS.K	△		Modèle BNTS, composantes temporelle et sociale désactivées
BNTS.KO	△		Modèle BNTS, composante sociale désactivée
BNTS.KS	△		Modèle BNTS, composante temporelle désactivée
Nestor	◊	∗	Notre modèle bayésien fondé sur un réseau d'inférence (Jabeur *et al.,* 2012)
isiFDL	△	∗	Modèle *MRF* avec apprentissage d'ordonnancements, 1er système dans le classement de TREC Microblog 2011 (Metzler *et al.,* 2011)
DFReeKLIM30	△	∗	Modèle basé sur la divergence de *Kullback-Leibler*, 2ème système dans le classement de TREC Microblog 2011(Amati *et al.,* 2011)
Disjunctive	○	∗	Modèle thématique disjonctif basé sur le système de RI *Lucene*
BM25	○		Modèle thématique d'*Okapi BM25*

△ Seuil à 30 *tweets*. ◊ Seuil automatique. ○ Aucun seuil. ∗ Résultats officiels.

Tableau 2 – Notation des différents modèles et configurations

5.2 Paramétrage du modèle

La figure 4.a présente l'impact de la fréquence des termes sur l'efficacité de recherche. Les précisions $p@30$ en fonction du paramètre β sont obtenues par la configuration BTNS.K. Pour $\beta = 0$, la fréquence est remplacée par une valeur binaire indiquant la simple présence d'un terme dans le *tweet*. Plus la valeur de β est grande, plus l'intervalle de transformation des fréquences est large et plus nous donnons d'importance aux fréquences élevées. La précision $p@30$ atteint une valeur maximale avec $\beta = 0.5$. Au-delà de cette valeur, on remarque une importante dégradation des performances, évaluée à -60% pour $\beta = 1$. Ces résultats ainsi que les analyses présentées dans la figure 3.a permettent de conclure que la fréquence élevée de certains termes de la requête est moins significative que la présence binaire de l'ensemble des termes.



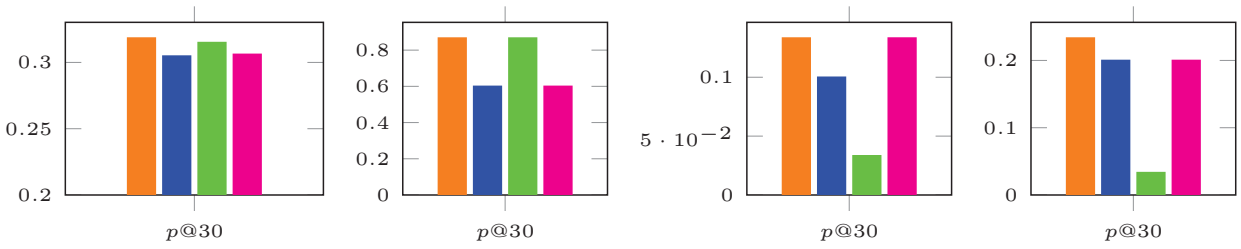
(a) Importance des fréquences (BNTS.K) (b) Taille de la fenêtre temporelle (BNTS.KO)

Figure 4 – Ajustements des paramètres β et Δt

La figure 4.b présente l'effet de la taille de fenêtre temporelle Δt sur l'efficacité de la recherche. Les précisions $p@30$ en fonction de Δt sont obtenues par la configuration BTNS.KO. Nous constatons que la précision $p@30$ atteint une valeur maximale à $\Delta t = 1j$. En dehors de l'intervalle $[4h, 30h]$, nous remarquons une dégradation des performances par rapport au modèle thématique BNTS.K qui correspond dans la figure au point d'abscisse $\Delta t = 0j$. Pour le reste des expérimentations, nous retenons les valeurs optimales de β et Δt qui permettent de maximiser la précision $p@30$: $\beta = 0.5$ et $\Delta t = 1j$.

5.3 Comparaison des facteurs de pertinence

La figure 5 permet de comparer l'impact des facteurs temps et importance sociale sur l'efficacité de recherche. Nous constatons de la première figure 5.a que la combinaison des 3 évidences donnée par la configuration BNTS, permet d'améliorer les résultats de la configuration BNTS.K basée seulement sur l'évidence thématique. En deuxième position, nous trouvons la configuration basée sur l'évidence thématique et temporelle BNTS.KO. La configuration basée sur l'évidence thématique et sociale BNTS.SO et la configuration BNTS.K présentent des résultats plus bas.



(a) Toutes les requêtes (b) MB001 (c) MB010 (d) MB027

Figure 5 – Comparaison des différentes configurations du modèle

Nous présentons dans les figures 5.b, 5.c et 5.d une analyse par requête spécifique. Ces figures montrent une variation plus importante entre les valeurs de la $p@30$. Pour la requête MB001 *"BBC World Service staff cuts"*, la plus haute valeur est obtenue par les deux configurations BNTS et BNTS.KO. Le facteur temps est donc un facteur primordial pour ce cette requête. Pour la requête MB010 *"Egyptian protesters attack museum"*, les deux configurations BNTS et BNTS.KS présentent les valeurs de précision les plus élevées. Dans ce cas, l'évidence sociale semble la plus appropriée pour répondre à cette requête. Finalement, les résultats obtenus par la configuration BNTS pour la requête MB027 *"reduce energy consumption"* dépassent toutes les autres configurations. Cela montre l'intérêt de la combinaison des trois sources d'évidence pour certaines requêtes de nature mixte. En conclusion, l'importance de chaque source d'évidence dépend de la nature de la requête et la motivation de recherche.

5.4 Évaluation de l'efficacité du modèle

Le tableau 3 présente les précisions obtenues par notre modèle BNTS et les autres approches similaires. Nous notons une amélioration de 57% de la précision $p@30$ par rapport à notre modèle précédent basé sur le réseau d'inférence bayésien Nestor. Comparativement aux deux premiers systèmes dans le classement final de TREC Microblog 2011 isiFDL et DFReeKLIM30, notre modèle présente des précisions plus basses. Cependant, nous notons un gain de 24% par rapport à la médiane des précisions $p@30$ des résultats officiels. Sur l'ensemble des 49 requêtes, 29 présentent des valeurs de $p@30$ qui dépassent les médianes correspondantes. 3 requêtes présentent des resultats similaires aux médianes. En comparant notre modèle aux *ba-*

selines BM25 et Disjunctive qui n'appliquent de seuil sur le nombre des résultats, nous constatons un gain de la MAP de 5% et de 13% respectivement. Par conséquent, notre technique de filtrage est donc capable de maintenir une valeur de MAP stable malgré la réduction de la liste des résultats.

	p@30			MAP	
isiFDL	0.4551	-30%		0.1923	-17%
DFReeKLIM30	0.4401	-28%		0.2348	-32%
BNTS	0.3184			0.1594	
Médiane	0.2575	24%		0.142	12%
Nestor	0.2027	57%	$*$	0.1305	22%
BM25	0.1136	180%	$***$	0.1517	5%
Disjunctive	0.0986	223%	$***$	0.1411	13%

Tableau 3 – Amélioration de la p@30 et de la MAP de notre modèle par rapport à celles des modèles de référence ($* : t.test < 0.05$; $*** : t.test < 0.001$)

6 Conclusion

Nous avons proposé dans cet article un modèle social pour la recherche de *tweets* basé sur les réseaux bayésiens de croyance. Ce modèle intègre trois sources d'évidence qui consistent en l'évidence thématique, l'évidence temporelle et l'évidence sociale. La pertinence thématique est estimée selon une fonction de fréquence tf modifiée. Ainsi, la présence de l'ensemble des termes de la requête est favorisée. La pertinence temporelle des *tweets* est déduite de la distribution des termes dans les fenêtres temporelles. Finalement, la pertinence sociale est estimée par l'autorité du blogueur correspondant dans le réseau social de *retweet*. L'évaluation expérimentale menée sur la collection TREC Microblog 2011 montre que la combinaison de ces différentes sources d'évidence permet de mieux évaluer la qualité des *tweets*. Les résultats obtenus dépassent la médiane des résultats officiels avec un gain de 24%.

En perspective, nous envisageons de détecter automatiquement la taille de la fenêtre temporelle selon la distribution temporelle des *tweets*. Nous envisageons également d'adapter la combinaison des différentes sources d'évidence selon la nature de la requête qui peut être thématique, sociale ou temporelle.

7 Bibliographie

Amati G., Amodeo G., Bianchi M., Celi A., Nicola C. D., Flammini M., Gaibisso C., Gambosi G., , Marcone G., « FUB, IASI-CNR, UNIVAQ at TREC 2011 », *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, 2011.

Cavnar W. B., Trenkle J. M., « N-Gram-Based Text Categorization », *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, p. 161-175, 1994.

Cha M., Haddadi H., Benevenuto F., Gummadi K., « Measuring User Influence in Twitter : The Million Follower Fallacy », *Proceedings of the 4th international AAAI Conference on Weblogs and Social Media*, ICWSM '10, 2010.

Choudhury M. D., Counts S., Czerwinski M., « Find Me the Right Content ! Diversity-Based Sampling of Social Media Spaces for Topic-Centric Search », *Proceedings of the 5th international AAAI Conference on Weblogs and Social Media*, ICWSM ’11, 2011.

Duan Y., Jiang L., Qin T., Zhou M., Shum H.-Y., « An empirical study on learning to rank of tweets », *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING ’10, p. 295-303, 2010.

Grinev M., Grineva M., Boldakov A., Novak L., Syssoev A., Lizorkin D., « Sifting micro-blogging stream for events of user interest », *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR ’09, ACM, New York, NY, USA, p. 837-837, 2009.

Jabeur L. B., Tamine L., Boughanem M., « Uprising microblogs : A Bayesian network retrieval model for tweet search », *Proceedings of the 2012 ACM Symposium on Applied Computing*, SAC ’11, 2012.

Java A., Song X., Finin T., Tseng B., « Why we twitter : understanding microblogging usage and communities », *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD ’07, 2007.

Kwak H., Lee C., Park H., Moon S., « What is Twitter, a social network or a news media ? », *Proceedings of the 19th international conference on World wide web*, WWW ’10, ACM, New York, NY, USA, p. 591-600, 2010.

Metzler D., Cai C., « USC/ISI at TREC 2011 : Microblog Track », *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, 2011.

Nagmoti R., Teredesai A., De Cock M., « Ranking Approaches for Microblog Search », *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT ’10, IEEE Computer Society, Washington, DC, USA, p. 153-157, 2010.

Ounis I., Macdonald C., Lin J., Soboroff I., « Overview of the TREC2011 Microblog Track », *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, 2011.

Pal A., Counts S., « Identifying topical authorities in microblogs », *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM ’11, ACM, New York, NY, USA, p. 45-54, 2011.

Sankaranarayanan J., Samet H., Teitler B. E., Lieberman M. D., Sperling J., « TwitterStand : news in tweets », *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS ’09, p. 42-51, 2009.

Silva I., Ribeiro-Neto B., Calado P., Moura E., Ziviani N., « Link-based and content-based evidential information in a belief network model », *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR ’00, ACM, New York, NY, USA, p. 96-103, 2000.

Teevan J., Ramage D., Morris M. R., « #TwitterSearch : a comparison of microblog search and web search », *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM ’11, ACM, New York, NY, USA, p. 35-44, 2011.

Weng J., Lim E.-P., Jiang J., He Q., « TwitterRank : finding topic-sensitive influential twitterers », *Proceedings of the third ACM international conference on Web search and data mining*, WSDM ’10, ACM, New York, NY, USA, p. 261-270, 2010.

Yardi S., Romero D. M., Schoenebeck G., Boyd D., « Detecting Spam in a Twitter Network », *First Monday*, 2010.