Tri des actualités sociales: État de l'art et Pistes de recherche

Sami Belkacem*, Kamel Boukhalfa*, Omar Boussaid**

*Laboratoire LSI, USTHB-Alger, Algérie - {s.belkacem,kboukhalfa}@usthb.dz

**Laboratoire ERIC, Université de Lyon, Lyon 2, France - omar.boussaid@univ-lyon2.fr

Résumé. En raison de la grande quantité d'informations (messages, articles, vidéos, musiques, images, etc.) produites et partagées sur les réseaux sociaux, les utilisateurs se retrouvent submergés d'informations générées chronologiquement dans leur fil d'actualité. De plus, la majorité des informations peuvent s'avérer non pertinentes. Le tri, par ordre de pertinence des actualités sociales, est proposé comme une solution pour aider les utilisateurs à consulter et interagir rapidement avec les informations susceptibles de les intéresser. Dans ce travail, nous étudions les approches existantes dans le domaine du tri des actualités sociales, et exposons leurs limites et quelques pistes de recherche selon plusieurs axes: les facteurs influençant la pertinence des informations, les modèles de prédiction de la pertinence, l'apprentissage et l'évaluation des modèles de prédiction, etc.

1 Introduction

Les réseaux sociaux occupent de plus en plus de place dans notre quotidien (Zhan et al., 2016). Les données manipulées sur ces réseaux sont connues pour leurs volumes qui peuvent atteindre des Pétaoctets (10^{15} octets) voire plus, leur hétérogénéité (messages, articles, vidéos, musiques, images, etc.), leur variété (pouvant provenir de différentes sources), et leur vélocité (arrivant en temps réel ou presque) (Xu et al., 2016). Ces caractéristiques ont fait que les technologies de gestion et de traitements classiques se trouvent dans l'incapacité de traiter ce type de données (Krishnan, 2013). Les données sociales constituent un des volets ayant contribué à l'apparition du concept de *big data* qui est défini par les 5 "V" : Volume, Variété, Vélocité, Valeur et Véracité (Lomotey et Deters, 2014). Sur ces réseaux, en raison de la grande quantité d'informations produites et partagées (Guy et al., 2011), les utilisateurs se retrouvent submergés d'informations générées chronologiquement dans leur fil d'actualité [1] (Berkovsky et al., 2012). Pour un utilisateur standard de *Facebook* [2] par exemple, 1500 nouvelles informations sont générées chaque jour dans son fil d'actualité [3]. En outre, plusieurs travaux de recherche ont montré que la majorité de ces informations sont considérées comme non pertinentes (Hong

1. Liste d'informations récentes (publications, posts, tweets, etc.) qui permet à un utilisateur de suivre l'actualité des membres de son réseau social.
2. www.facebook.com
3. www.slate.fr/story/112681/qui-controle-ce-qui-apparait-sur-votre-fil-facebook

et al., 2012). Par exemple, Paek et al. (2010) ont demandé à 24 utilisateurs de *Facebook* d'associer des scores de pertinence aux informations de leur fil d'actualité. La moyenne globale de ces scores était proche de 0. De ce fait, il devient difficile pour les utilisateurs de consulter et d'interagir (cliquer, commenter, aimer, partager, etc.) rapidement avec les informations pertinentes (Lakkaraju et al., 2011), notamment pour les utilisateurs ayant un nombre important de relations sociales (Pan et al., 2013; Kuang et al., 2016). *Facebook* affirme qu'un utilisateur standard reste susceptible de manquer une partie des informations pertinentes même si ce dernier passe une moyenne de 55 minutes par jour sur le réseau social (Paek et al., 2010).

En se basant sur la prédiction d'un score de pertinence entre un utilisateur et une nouvelle information non consultée dans son fil d'actualité, des travaux de recherche ont proposé des approches pour trier et afficher les actualités sociales par ordre décroissant de pertinence (Berkovsky et Freyne, 2015). Un état de l'art couvrant 3 différents axes a été effectué dans Berkovsky et Freyne (2015). Le premier s'intéresse aux utilisateurs générateurs d'informations, à ceux qui consultent et bénéficient de ces informations et aux liens entre les deux. Le deuxième axe porte sur le contenu de l'information. Le troisième axe concerne la structure du réseau social sous-jacent des utilisateurs et des informations qu'ils génèrent. Cependant, à notre avis, le travail proposé ne couvre pas 4 autres axes importants : les facteurs influençant la pertinence des informations, les modèles de prédiction de la pertinence, l'apprentissage et l'évaluation des modèles de prédiction et les réseaux sociaux cibles. En considérant ces 4 axes, et dans le souci de compléter l'état de l'art effectué dans Berkovsky et Freyne (2015), l'objectif du présent travail est de faire un état de l'art sur les approches existantes dans le domaine du tri des actualités sociales, afin de montrer leurs avantages, leurs limites et identifier des pistes de recherche. Cela nous amène à formuler les questions suivantes : d'une part, quels sont les facteurs influençant la pertinence des informations pour les utilisateurs ? D'autre part, quel modèle utiliser pour prédire la pertinence d'une nouvelle information non consultée dans le fil d'actualité d'un utilisateur à partir de ces facteurs ? Également, étant donné que les utilisateurs ne donnent pas explicitement les scores de pertinence relatifs aux informations de leur fil d'actualité (Berkovsky et Freyne, 2015), comment faire l'apprentissage et l'évaluation du modèle de prédiction en question ? Enfin, quels sont les réseaux sociaux ciblés par les travaux existants ? Étudier ces questions est l'objet de notre papier. Nous notons que le présent papier est un *position paper*. Il consiste en un état de l'art et donc, n'inclut pas d'expérimentations. Toutefois, nous présenterons une synthèse des travaux étudiés.

Le papier est structuré comme suit : la section 2 présente des généralités sur le tri des actualités sociales, la section 3 présente une étude des travaux effectués dans ce domaine et expose leurs limites ainsi que quelques pistes de recherche, la section 4 présente la conclusion et nos perspectives.

2 Tri des actualités sociales

Dans cette section, nous présentons des généralités sur le tri des actualités sociales incluant la définition des fils d'actualité, la présentation de statistiques qui confirment la nécessité du tri et la définition du tri des actualités dans ces fils.

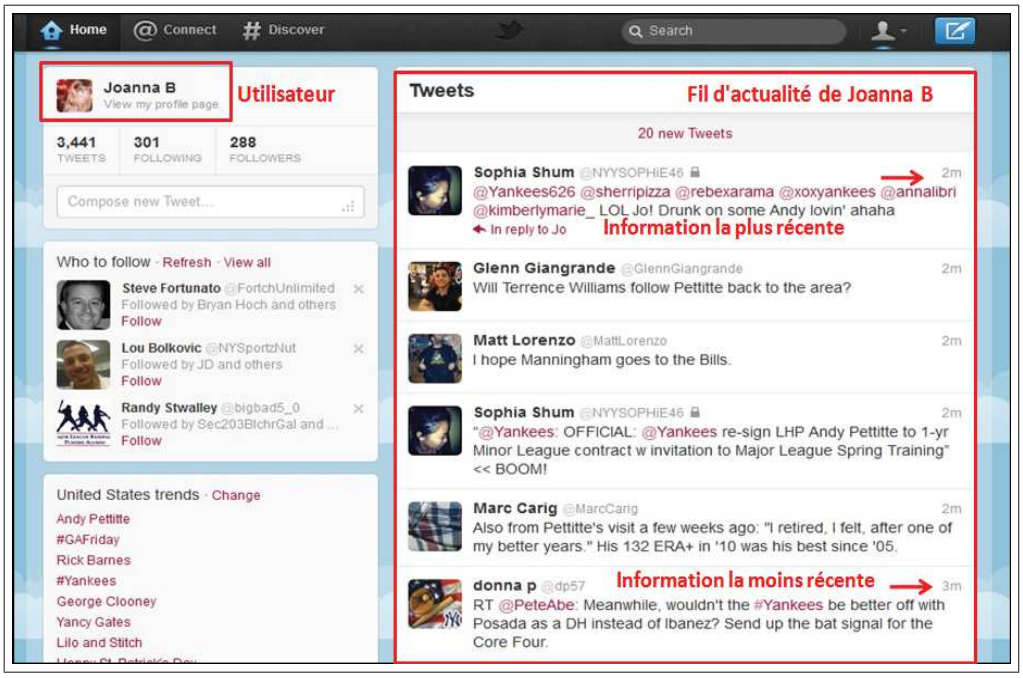


FIG. 1: Fil d'actualité d'un utilisateur sur *Twitter*.

2.1 Définition du fil d'actualité

Selon Rader et Gray (2015), le fil d'actualité d'un utilisateur (voir Figure 1) est une liste d'informations (publications, posts, tweets, etc.) qui lui permet de suivre l'actualité des membres de son réseau social. Il comprend des messages textuels, articles, vidéos, musiques, images, etc. (Freyne et al., 2010). Sur un réseau social grand public comme *Facebook* ou *Twitter*[4], le fil d'actualité d'un utilisateur est constitué d'actualités (voir Figure 2) relatives à ses amis, membres de sa famille, pages auxquelles il s'est abonné, etc. (Ester, 2013). Par contre, sur un réseau social professionnel et/ou académique comme ResearchGate[5] ou *LinkedIn*[6], le fil d'actualité d'un utilisateur est constitué d'actualités relatives à ses contacts, collègues, camarades, etc. (Ester, 2013).

Sur la plupart des réseaux sociaux, et comme le montre l'exemple de la Figure 1, le fil d'actualité est affiché par ordre chronologique, de l'information la plus récente jusqu'à la moins récente (Berkovsky et al., 2012). L'inconvénient avec ce fil chronologique est que l'utilisateur doit consulter et parcourir toutes les informations présentes dans son fil pour être sûr de ne manquer celle qui pourrait être pertinente (Kuang et al., 2016).

4. www.twitter.com
5. www.researchgate.net
6. www.linkedin.com

Généralement, une actualité possède (voir Figure 2 (a)) : (1) un auteur qui a généré l'information ; (2) un ensemble d'utilisateurs bénéficiaires qui peuvent consulter et interagir (cliquer, commenter, aimer, partager, etc.) avec l'information ; (3) un contenu textuel et/ou multimédia ; (4) une date de publication sur le réseau social (5) un espace pour effectuer des actions sur l'information : enregistrer, activer les notifications, signaler, masquer, etc. ; (6) des tags, ou marqueurs de métadonnées précédés d'hashtag "#", qui décrivent la thématique de l'information ; (7) des mentions, ou noms précédés d'arobase "@", qui représentent des liens vers d'autres utilisateurs du réseau social ; et enfin (8) des URLs vers des sites internet ou des articles (Freyne et al., 2010; Berkovsky et Freyne, 2015).

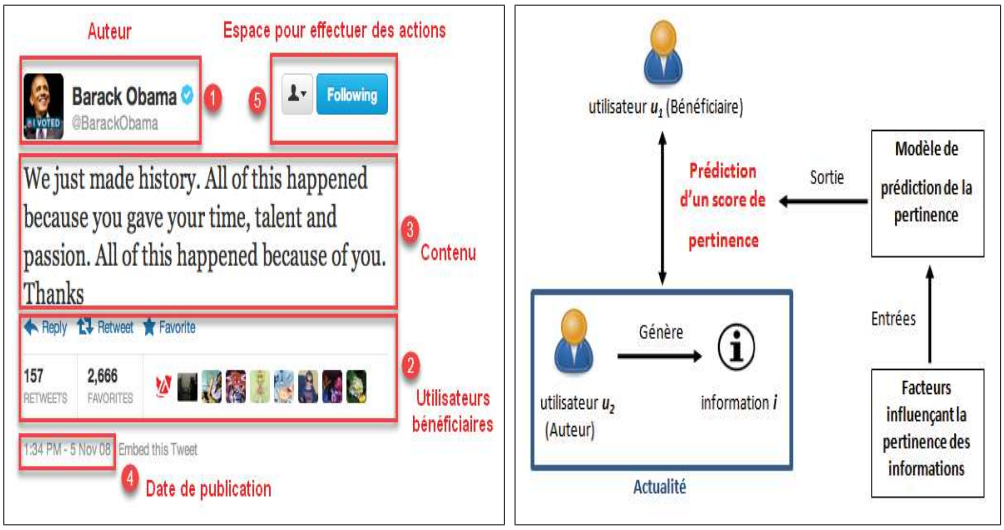
2.2 Statistiques sur les fils d'actualité

Au vu de certaines statistiques sur le volume important d'informations et de leur non-pertinence, la nécessité du tri est plus que recommandée.

Volume important. Pour un utilisateur standard de *Facebook*, 1500 nouvelles informations sont générées chaque jour dans son fil d'actualité. Cela peut monter jusqu'à 10000 informations pour les utilisateurs ayant plusieurs centaines de relations sociales[3]. Sur le réseau social *Instagram*[7], en raison du volume important d'informations, les utilisateurs ne voient que 30% des informations de leur fil d'actualité[8]. Dans Bontcheva et al. (2013), 587 utilisateurs de *Twitter* ont répondu à un questionnaire. Les résultats montrent que 66,3% des utilisateurs ont parfois le sentiment qu'ils ne peuvent pas suivre le grand volume d'informations dans leur fil d'actualité. Dans Ramage et al. (2010), suite à un sondage auprès de 56 utilisateurs de *Twitter*, les auteurs rapportent que les utilisateurs ont trop d'informations dans leur fil d'actualité.

Non-pertinence. Pour montrer la non-pertinence des informations, Paek et al. (2010) ont demandé à 24 utilisateurs de *Facebook* d'associer des scores de pertinence aux informations de leur fil d'actualité. De même, Alonso et al. (2013) ont demandé à 5 utilisateurs de *Twitter* d'associer des scores de pertinence à plus de 2000 tweets. Dans les deux cas, la moyenne globale des scores de pertinence obtenus était proche de 0. Dans Bontcheva et al. (2013), 587 utilisateurs de *Twitter* ont répondu à un questionnaire. Les résultats montrent que 70.4% des utilisateurs ont des difficultés pour trouver les tweets pertinents dans leur fil d'actualité. Dans Ramage et al. (2010), suite à un sondage auprès de 56 utilisateurs de *Twitter*, les auteurs rapportent que les utilisateurs perdent les tweets les plus pertinents dans un flux de milliers de tweets de moindre utilité. Sur *LinkedIn*, Agarwal et al. (2014) affirment que le fil d'actualité chronologique conduit à un fil récent mais pas nécessairement pertinent. Les auteurs ont effectué un test en ligne comparant le fil chronologique avec un fil basé sur la pertinence, et ont trouvé le taux de clics[9] du fil basé sur la pertinence 43% supérieur au fil chronologique.

7. www.instagram.com
8. www.presse-citron.net/instagram-deploie-son-fil-dactualite-non-chronologique
9. Le taux de clics est un rapport entre le nombre de clics qu'un élément reçoit et le nombre d'affichages de celui-ci.



(a) Actualité sur *Twitter*.

(b) Pertinence d'une actualité.

FIG. 2: Actualité sur un réseau social.

2.3 Définition du tri des actualités sociales

Selon Shen et al. (2013), le tri des actualités sociales consiste à trier, par ordre décroissant de pertinence, les informations du fil d'actualité de chaque utilisateur. Le tri se fait de telle sorte que les informations les plus pertinentes se retrouvent en haut du fil d'actualité et les moins pertinentes en bas (Agarwal et al., 2014). Nous notons que d'autres termes ont été utilisés dans la littérature pour faire référence au *"tri"* des actualités sociales comme : recommandation [10], personnalisation, classement, réorganisation, etc.

Le tri des actualités sociales peut être considéré soit comme une top-k recommandation ou un problème de reclassement (Berkovsky et Freyne, 2015). Les techniques utilisées pour classer et trier les informations sont basées sur un modèle de prédiction, qui utilise en entrée un ensemble de facteurs influençant la pertinence des informations, pour produire en sortie un score de pertinence entre un utilisateur bénéficiaire u_1 et une nouvelle information i non consultée dans son fil d'actualité, générée par un auteur u_2 (Berkovsky et al., 2012) (voir Figure 2 (b)).

3 Travaux de recherche sur le tri des actualités sociales

Dans le domaine du tri des actualités sociales, plusieurs travaux ont été effectués tant dans la communauté scientifique que dans la communauté industrielle. Dans cette section, nous présentons ces travaux selon les communautés.

10. La recommandation vise, à partir de connaissances sur un utilisateur, à lui faciliter l'accès aux contenus pertinents (messages, articles, vidéos, musiques, images, etc.) dans un catalogue trop vaste.

3.1 Travaux dans la communauté industrielle

Facebook, *Twitter* et *LinkedIn*, ont fourni des efforts dans le tri des actualités sociales (Berkovsky et Freyne, 2015). Cependant, les approches existantes n'ont pas été divulguées en raison de la sensibilité commerciale et de la compétitivité entre les entreprises (Berkovsky et Freyne, 2015). De plus, ces dernières affirment que leurs algorithmes affichent plusieurs limites[3] (Agarwal et al., 2014, 2015). Toutefois, si *Facebook*, *Twitter* et *LinkedIn* s'intéressent au tri des actualités sociales, ce n'est pas uniquement pour satisfaire et fidéliser les utilisateurs, mais également pour booster le taux d'interaction en privilégiant les informations pertinentes susceptibles de faire interagir les utilisateurs[3]. En effet, il s'avère que les interactions des utilisateurs constituent le carburant économique de ces entreprises[3].

Will Oremus, journaliste à *Slate.com*, a pu rencontrer en 2016 une équipe de *Facebook* en charge du fil d'actualité[3]. Il affirme que l'algorithme de tri utilisé par *Facebook* combine des centaines de facteurs pour prédire la pertinence des actualités sociales. Cependant, il souligne que l'algorithme reste susceptible de proposer des informations que les utilisateurs trouvent non pertinentes. Effectivement, selon le journaliste, le réseau social a organisé un test auprès de certains utilisateurs, en leur montrant la première information de leur fil d'actualité à côté d'une autre, moins pertinente, afin qu'ils choisissent celles qu'ils préfèreraient lire. Les résultats ont montré que le tri effectué par l'algorithme correspond parfois aux préférences de l'utilisateur. Lorsque les résultats ne correspondent pas, l'équipe *Facebook* affirme que cela indique un point à améliorer.

3.2 Travaux dans la communauté académique

Nous avons recensé 17 travaux de recherche dans la communauté académique (Paek et al., 2010; Freyne et al., 2010; Berkovsky et al., 2012; Uysal et Croft, 2011; Feng et Wang, 2013; Shen et al., 2013; Chen et al., 2012; Lakkaraju et al., 2011; Kuang et al., 2016; Agarwal et al., 2014, 2015; Guy et al., 2011; Hong et al., 2012; Yan et al., 2012; Bourke et al., 2013; Soh et al., 2013; Pan et al., 2013). Par manque d'espace, nous considérons, dans le cadre du présent papier, les 11 travaux les plus représentatifs (11 premiers travaux cités dans la liste ci-dessus), soit 65% de tous les travaux recensés.

Dans le but de prédire la pertinence des actualités sur *Facebook*, Paek et al. (2010) et Lakkaraju et al. (2011) ont proposé des modèles personnalisés [11] qui exploitent 3 types de facteurs de prédiction de la pertinence : (1) la force de la relation sociale entre le bénéficiaire et l'auteur ; (2) la pertinence du contenu de l'information pour les intérêts du bénéficiaire ; et (3) la qualité de l'information. Dans Paek et al. (2010), le modèle d'apprentissage supervisé de classification [12] proposé se base sur les machines à vecteurs de support (SVM) (Cortes et Vapnik, 1995). Pour obtenir les données d'apprentissage et d'évaluation du modèle de prédiction, les auteurs ont demandé à 24 utilisateurs d'associer explicitement des scores de pertinence aux informations de leur fil d'actualité. La précision moyenne du modèle était de 69,7%. Tandis que dans Lakkaraju et al. (2011), le modèle de filtrage collaboratif proposé se base sur les facteurs latents de prédiction. Pour obtenir les données d'apprentissage et d'évaluation du modèle de

11. Un modèle personnalisé est un modèle qui prend en considération les préférences de chaque utilisateur.
12. Les scores de pertinences à prédire appartiennent à un ensemble fini de valeurs.

prédiction, les interactions des utilisateurs en termes de commentaires ont été utilisées comme indicateurs implicites de la pertinence des informations. La précision moyenne du modèle était de 61.11%.

Afin d'associer des scores de pertinence aux actualités : sur le réseau social *SocialBlue* dans Freyne et al. (2010), une communauté en ligne liée à la santé dans Berkovsky et al. (2012) et le réseau social chinois *Sina Weibo* [13] dans Kuang et al. (2016), tous ces auteurs ont proposé des modèles non personnalisés de prédiction basés sur des combinaisons linéaires pondérées avec des poids statiques. Dans Freyne et al. (2010) et Berkovsky et al. (2012), les modèles proposés exploitent 2 facteurs de prédiction de la pertinence : (1) la force de la relation sociale entre le bénéficiaire u_1 et l'auteur u_2 ; et (2) la fréquence d'interaction du bénéficiaire avec des informations similaires à l'information i (voir section 2.3). Dans Freyne et al. (2010), la précision moyenne du modèle était de 75% et les résultats de l'évaluation ont montré que les informations triées ont réussi à attirer plus d'attention que celles non triées. Dans Berkovsky et al. (2012), la précision moyenne a été améliorée de 10% par rapport aux résultats du modèle chronologique. Dans les deux précédents travaux, afin d'obtenir les données d'évaluation du modèle de prédiction, les consultations et interactions des utilisateurs (authentification, commentaires, messages, etc.) sur le réseau social, ont été utilisées comme indicateurs implicites de la pertinence des informations. Dans Kuang et al. (2016), le modèle proposé exploite 3 types de facteurs de prédiction de la pertinence : (1) la force de la relation sociale entre l'auteur et le bénéficiaire ; (2) la pertinence du contenu de l'information (et de ses tags, mentions et URLs) pour les intérêts du bénéficiaire ; et (3) la qualité de l'information. Pour obtenir les données d'évaluation du modèle de prédiction, les auteurs ont demandé à 1048 utilisateurs d'associer explicitement des valeurs booléennes (Vrai pour pertinent et Faux pour non pertinent) aux informations de leur fil d'actualité. La précision moyenne du modèle était de 75% et a été améliorée de 57% par rapport aux résultats du modèle chronologique.

Dans l'intention de trier les tweets par ordre de pertinence sur *Twitter*, Uysal et Croft (2011) et Feng et Wang (2013) ont proposé des modèles personnalisés qui prédisent la probabilité qu'un bénéficiaire retweete un tweet présent dans son fil d'actualité. Dans Uysal et Croft (2011), le modèle proposé se base sur un algorithme ascendant coordonné et exploite 5 types de facteurs de prédiction de la pertinence : (1) la force de la relation sociale entre l'auteur et le bénéficiaire ; (2) la pertinence du contenu de l'information (et de ses tags, mentions et URLs) pour les intérêts du bénéficiaire ; (3) la qualité du tweet ; (4) l'autorité de l'auteur, et (5) l'activité de l'auteur sur le réseau social. La précision moyenne du modèle était de 72%. Tandis que dans Feng et Wang (2013), le modèle proposé se base sur la factorisation de matrices et exploite 5 types de facteurs de prédiction de la pertinence : (1) la force de la relation sociale entre l'auteur et le bénéficiaire ; (2) la pertinence du contenu de l'information (et de ses tags, mentions et URLs) pour les intérêts du bénéficiaire ; (3) la qualité du tweet ; (4) l'autorité de l'auteur ; et (5) les propriétés du bénéficiaire (préférences pour les différentes thématiques, ancienneté, probabilité de retweeter, date du dernier retweet effectué, etc.). La précision moyenne du modèle combiné était de 42%. Dans les deux précédents travaux, pour obtenir les données d'apprentissage et d'évaluation du modèle de prédiction, les interactions des utilisateurs, en termes de retweets ont été utilisées comme indicateurs implicites de la pertinence des tweets.

13. www.weibo.com

Afin de recommander des tweets pertinents aux utilisateurs de *Twitter*, Shen et al. (2013) et Chen et al. (2012) ont proposé des modèles personnalisés de prédiction qui exploitent 4 types de facteurs de prédiction : (1) la force de la relation sociale entre l'auteur et le bénéficiaire ; (2) la pertinence du contenu de l'information (et de ses tags, mentions et URLs) pour les intérêts du bénéficiaire ; (3) la qualité du tweet ; et (4) l'autorité de l'auteur. Dans Shen et al. (2013), les auteurs ont proposé un modèle personnalisé d'apprentissage supervisé. Le modèle de régression [14] proposé se base sur le *Gradient Boosting* (Zheng et al., 2007). Pour obtenir les données d'apprentissage et d'évaluation du modèle de prédiction, les interactions des utilisateurs, en termes de tweets, retweets et réponses ont été utilisées comme indicateurs implicites de la pertinence des tweets. Le modèle proposé a connu un gain de précision moyenne de 34.5% comparé au modèle chronologique. Tandis que dans Chen et al. (2012), les auteurs ont proposé un modèle de filtrage collaboratif basé sur les facteurs latents de prédiction. Pour obtenir les données d'apprentissage et d'évaluation du modèle de prédiction, les interactions des utilisateurs en termes de retweets ont été utilisées comme indicateurs implicites de la pertinence des tweets. La précision moyenne du modèle était de 76% et les résultats de l'évaluation ont montré que les informations triées ont réussi à attirer plus d'attention que les informations non triées.

Afin d'associer des scores de pertinence aux actualités sociales sur *LinkedIn*, Agarwal et al. (2014) et Agarwal et al. (2015) ont proposé des modèles personnalisés basés sur la régression logistique. Les modèles proposés prédisent la probabilité qu'un utilisateur clic sur une information issue de son fil d'actualité. Dans Agarwal et al. (2014), le modèle proposé exploite 3 types de facteurs de prédiction de la pertinence : (1) la force de la relation sociale entre le bénéficiaire et l'auteur ; (2) la qualité de l'information ; et (3) la diversité du fil d'actualité en privilégiant la diversité des informations et leurs auteurs. Dans Agarwal et al. (2015), le modèle proposé exploite 3 types de facteurs de prédiction de la pertinence : (1) la force de la relation sociale entre le bénéficiaire et l'auteur ; (2) la fréquence d'interaction du bénéficiaire avec des informations similaires à l'information i ; et (3) la fréquence d'interaction du bénéficiaire avec des informations similaires à l'information i générées par l'auteur u_2. Dans les deux précédents travaux, pour obtenir les données d'apprentissage et d'évaluation du modèle de prédiction, les interactions des utilisateurs en termes de clics ont été utilisées comme indicateurs implicites de la pertinence des informations. Les résultats des évaluations ont montré que le taux de clics des modèles proposés a été amélioré, comparé au modèle chronologique.

3.3 Analyse des travaux de recherche

En se basant sur les questions posées dans la section 1, notre analyse des travaux de recherche existants s'articule autour de 4 axes : les facteurs influençant la pertinence des informations, les modèles de prédiction de la pertinence, l'apprentissage et l'évaluation des modèles de prédiction et les réseaux sociaux cibles. Le Tableau 1 résume notre analyse. Les cellules contenant le symbole "x" indiquent que le critère correspondant a été considéré dans le travail et les cellules vides indiquent qu'il ne l'a pas été. Par manque d'espace, nous annotons les travaux de recherche dans le tableau comme suit : *'A'* pour (Paek et al., 2010), *'B'* pour

14. Les scores de pertinences à prédire sont des valeurs dans un ensemble continu de réels.

(Lakkaraju et al., 2011), *'C'* pour (Freyne et al., 2010), *'D'* pour (Berkovsky et al., 2012), *'E'* pour (Uysal et Croft, 2011), *'F'* pour (Feng et Wang, 2013), *'G'* pour (Shen et al., 2013), *'H'* pour (Chen et al., 2012), *'I'* pour (Kuang et al., 2016), *'J'* pour (Agarwal et al., 2014) et *'K'* pour (Agarwal et al., 2015).

Facteurs influençant la pertinence des informations. Nous avons noté la prédominance des facteurs suivants : (1) ceux mesurant la pertinence du contenu de l'information, de ses tags, mentions et URLs pour les intérêts du bénéficiaire : similarité sémantique, score TF-IDF, produit scalaire, etc. Ces derniers sont des prédicteurs directs pour savoir si l'information sera pertinente pour le bénéficiaire (Paek et al., 2010; Kuang et al., 2016) ; (2) les facteurs mesurant la force de la relation sociale entre le bénéficiaire et l'auteur : fréquence des interactions sociales, date de la dernière interaction, nombre de contacts en commun, etc. Gilbert et Karahalios (2009) ont été pionniers dans ce domaine de recherche. L'hypothèse est que l'information peut être pertinente pour le bénéficiaire s'il a une forte relation sociale avec l'auteur (Agarwal et al., 2014, 2015) ; (3) les facteurs mesurant l'autorité et l'influence de l'auteur sur le réseau social : nombre d'abonnés, certification du compte, ancienneté, etc. L'hypothèse est que l'information peut être pertinente pour le bénéficiaire si l'auteur a une grande autorité et une grande influence sur le réseau social (Uysal et Croft, 2011; Feng et Wang, 2013) ; et (4) les facteurs mesurant la qualité de l'information : fraicheur, nombre de personnes ayant interagi avec l'information, présence d'images, de vidéos, d'URLs, d'hashtags, etc. L'hypothèse est que l'information peut être pertinente pour le bénéficiaire si elle est de bonne qualité (Chen et al., 2012; Shen et al., 2013).

Modèles de prédiction de la pertinence. Nous avons noté la prédominance des modèles d'apprentissage supervisé et des modèles mathématiques.

1. Modèles d'apprentissage supervisé : modèles d'apprentissage automatique qui généralisent pour des entrées inconnues ce qu'ils ont pu "apprendre" grâce aux exemples déjà traitées et validées dans la base d'apprentissage (Mohri et al., 2012). Nous avons recensé un modèle de classification avec les machines à vecteurs de support (SVM) (Paek et al., 2010) et un modèle de régression avec *Gradient Boosting* (Shen et al., 2013) ;
2. Modèles mathématiques : modèles qui utilisent des outils, équations et concepts mathématiques pour modéliser et prédire des comportements (Bender, 2012). Nous avons recensé des modèles statistiques basés sur les facteurs latents de prédiction (Lakkaraju et al., 2011; Chen et al., 2012) et la régression logistique (Agarwal et al., 2014, 2015), des modèles d'algèbre linéaire basés sur des combinaisons linéaires pondérées (Freyne et al., 2010; Berkovsky et al., 2012; Kuang et al., 2016) et la factorisation de matrices (Feng et Wang, 2013) et un modèle d'optimisation mathématique basé sur un algorithme ascendant coordonné (Uysal et Croft, 2011).

En effet, ces modèles sont adéquats pour le problème de prédiction des scores de pertinence car ils font partie des techniques de l'analyse prédictive dont le but est d'analyser des faits présents et passés pour faire des hypothèses prédictives sur des événements futurs (Nyce et CPCU, 2007; Eckerson, 2007). Toutefois, dans la littérature, à notre connaissance, aucune comparaison n'a été effectuée entre ces modèles pour déterminer ceux les plus précis.

Apprentissage et évaluation des modèles de prédiction. Nous avons recensé 2 méthodes d'apprentissage et d'évaluation des modèles de prédiction : (1) une méthode implicite selon l'hypothèse que l'interaction (cliquer, commenter, aimer, partager, etc.) d'un utilisateur avec une information implique sa pertinence pour ce dernier (Shen et al., 2013; Agarwal et al., 2014, 2015). Nous notons la forte prédominance de cette méthode ; et (2) une méthode explicite en demandant aux utilisateurs du réseau social d'associer, via des outils développés, des scores et des valeurs de pertinences aux informations de leur fil d'actualité (Paek et al., 2010; Kuang et al., 2016).

Réseaux sociaux cibles. Nous avons noté la prévalence de *Twitter* comme réseau social ciblé par les travaux de recherche existants (Uysal et Croft, 2011; Feng et Wang, 2013; Shen et al., 2013; Chen et al., 2012). Cela se justifie par : le flux important de tweets rencontrés par les utilisateurs (Kuang et al., 2016), la non-pertinence d'un nombre important de tweets (Alonso et al., 2013), la disponibilité des données (les profils des utilisateurs, les réseaux sociaux des utilisateurs et les tweets sont publiquement visibles par défaut) et la disponibilité de l'API pour la collecte de ces derniers (Berkovsky et Freyne, 2015).

3.4 Limites et pistes de recherche

En se basant sur les questions posées dans la section 1, et suite à l'analyse des travaux de recherche existants, nous avons constaté certaines limites et identifié quelques pistes de recherche que nous classons selon 4 axes : les facteurs influençant la pertinence des informations, les modèles de prédiction de la pertinence, l'apprentissage et l'évaluation des modèles de prédiction et les réseaux sociaux cibles.

Facteurs influençant la pertinence des informations. Nous avons recensé un facteur non considéré dans la littérature, mais qui pourrait être susceptible d'influencer la pertinence des informations. Il s'agit de la similarité entre les informations (leur contenu, leurs tags et leurs URLs) générées par le bénéficiaire et l'information i : similarité sémantique, score TF-IDF, produit scalaire, etc. L'hypothèse est que l'information i peut être pertinente pour le bénéficiaire s'il elle est similaire aux informations qu'il a l'habitude de générer. L'avantage de ce facteur est qu'il permet de cerner les centres d'intérêt et les préférences du bénéficiaire même si celui-ci n'interagit pas avec les informations de son fil d'actualité. De plus, étant donné que la précision du modèle de prédiction dépend des facteurs en entrée (Berkovsky et al., 2012), il serait judicieux d'effectuer un sondage auprès des utilisateurs afin de valider les facteurs déjà considérés dans la littérature et découvrir d'autres facteurs non encore pris en charge.

Modèles de prédiction de la pertinence. Nous avons recensé 2 travaux qui utilisent des combinaisons linéaires pondérées avec des poids statiques, non personnalisés, pour tous les utilisateurs (Freyne et al., 2010; Berkovsky et al., 2012; Kuang et al., 2016). Cependant, les préférences de ces derniers sont différentes (Berkovsky et Freyne, 2015). De ce fait, les actualités devraient être triées d'une manière personnalisée en fonction des préférences de chaque utilisateur. De plus, à partir des évaluations effectuées dans la littérature, il serait intéressant de faire une étude comparative des mesures d'évaluations utilisées et des résultats obtenus afin de déterminer ceux les plus précis. Aussi, par rapport aux méthodes d'apprentissage supervisé

Travaux de recherche				A	B	C	D	E	F	G	H	I	J	K
Facteurs influençant la pertinence des informations	Force de la relation sociale entre l'auteur et le bénéficiaire			x	x	x	x	x	x	x	x	x	x	x
	Pertinence du contenu de l'information, de ses tags, mentions et URLs pour les intérêts du bénéficiaire			x	x			x	x	x	x	x		
	Qualité de l'information			x	x			x	x	x	x	x	x	
	Autorité de l'auteur							x	x	x	x			
	Activité de l'auteur							x						
	Fréquence d'interaction du bénéficiaire avec des informations similaires à l'information *i*					x	x							x
	Propriétés du bénéficiaire								x					
	Diversité du fil d'actualité												x	
	Fréquence d'interaction du bénéficiaire avec des informations similaires à l'information *i* générées par l'auteur u_2													x
Modèles de prédiction de la pertinence	Apprentissage supervisé	classification avec SVM		x										
		régression avec *Gradient Boosting*								x				
	Mathématiques	Statistiques	facteurs latents de prédiction		x						x			
			régression logistique										x	x
		Algèbre linéaire	combinaisons linéaires pondérées			x	x					x		
			factorisation de matrices						x					
		Optimisation mathématique	algorithme ascendant coordonné					x						
Apprentissage et évaluation	Implicite				x	x	x	x	x	x	x		x	x
	Explicite			x								x		
Réseaux sociaux cibles	*Twitter*							x	x	x	x			
	Facebook			x	x									
	SocialBlue					x								
	Communauté en ligne liée à la santé						x							
	Sina Weibo											x		
	LinkedIn												x	x

TAB. 1: Résumé de l'analyse des travaux de recherche.

(Paek et al., 2010; Shen et al., 2013), afin de prédire un score amélioré de pertinence pour chaque couple utilisateur-information, et pour mieux trier les informations, nous estimons que le problème de prédiction des scores de pertinence devrait être considéré comme un problème de régression (Shen et al., 2013) et non de classification (Paek et al., 2010). En effet, on parle de régression lorsque la sortie que l'on cherche à estimer est une valeur dans un ensemble continu de réels (Bishop, 2006). Par contre, on parle de problèmes de classification lorsque l'ensemble des valeurs de sortie est fini (Bishop, 2006). En outre, il serait judicieux d'utiliser les réseaux de neurones artificiels (McCulloch et Pitts, 1943) et/ou Adaboost (Freund et Schapire, 1995), connus pour donner de bons résultats (Schwenk et Bengio, 2000).

Apprentissage et évaluation des modèles de prédiction. La méthode implicite utilisée dans la plupart des travaux comporte plusieurs limites. En effet, une interaction avec une information n'est pas toujours synonyme de pertinence (Uysal et Croft, 2011; Shen et al., 2013). Par exemple, un utilisateur peut interagir pour exprimer son mécontentement. Dans d'autres cas, une absence d'interaction n'est pas toujours synonyme de non-pertinence (Uysal et Croft, 2011; Shen et al., 2013). Par exemple, un utilisateur peut trouver une information pertinente et choisir délibérément de ne pas interagir avec elle. La méthode explicite utilisée dans Paek et al. (2010) et Kuang et al. (2016) comporte aussi plusieurs limites. En effet, elle n'est pas incluse sur les réseaux sociaux (des outils ont été développés pour avoir les *feedbacks* des utilisateurs) et elle est contraignante, puisqu'elle demande aux utilisateurs d'associer des scores et des valeurs de pertinence à un nombre important d'informations. De ce fait, il serait judicieux d'effectuer un sondage auprès des utilisateurs afin de déterminer les indicateurs qui permettent d'approximer les scores de pertinence relatifs aux informations de leur fil d'actualité. En outre, dans le but d'extraire des indicateurs décisionnels, et des indicateurs de pertinence, il serait intéressant d'entreposer les données sociales pour effectuer des analyses en ligne et bénéficier des outils existants en informatique décisionnelle. Plusieurs questions restent ouvertes dans ce contexte notamment sur l'approche d'entreposage, le formalisme utilisé ainsi que les techniques de stockage des données (Hannachi et al., 2015; Oukid et al., 2016).

Réseaux sociaux cibles Étant donné que la plupart des travaux de recherche ont ciblé *Twitter* (Uysal et Croft, 2011; Feng et Wang, 2013; Shen et al., 2013; Chen et al., 2012), il serait intéressant d'exploiter d'autres réseaux sociaux qui disposent de fil d'actualité, mais qui n'ont pas été traitées dans les travaux de recherche, comme : *Instagram*, *Flickr* [15], *Pinterest* [16], etc.

4 Conclusion

Dans ce travail, nous avons d'abord présenté des généralités sur le tri des actualités sociales. Nous avons ensuite présenté une étude de l'état de l'art sur les approches existantes dans ce domaine. À l'issue de cette étude, il apparaît que les chercheurs accordent de plus en plus d'intérêt au problème du tri des actualités sociales. Plusieurs approches ont été proposées dans la littérature. Cependant, vu les limites recensées (voir section 3.4), des efforts doivent être faits pour améliorer le tri des actualités sociales. Les efforts à fournir s'articulent autour

15. www.flickr.com
16. www.pinterest.com

de 4 axes : les facteurs influençant la pertinence des informations, les modèles de prédiction de la pertinence, l'apprentissage et l'évaluation des modèles de prédiction et les réseaux sociaux cibles. Nos perspectives de recherche concernent les 3 premiers axes.

En plus des facteurs de prédiction considérés dans la littérature, nous comptons inclure la similarité entre les informations générées par le bénéficiaire et l'information *i*. De plus, afin de mieux trier les informations, nous projetons d'utiliser un modèle personnalisé de prédiction de la pertinence. En outre, dans le but d'extraire des indicateurs décisionnels, et des indicateurs de pertinence, nous planifions d'entreposer les données sociales pour effectuer des analyses en ligne et bénéficier des outils existants en informatique décisionnelle.

Références

Agarwal, D., B.-C. Chen, R. Gupta, J. Hartman, Q. He, A. Iyer, S. Kolar, Y. Ma, P. Shivaswamy, A. Singh, et others (2014). Activity ranking in LinkedIn feed. In *Proc. of the 20th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pp. 1603–1612.

Agarwal, D., B.-C. Chen, Q. He, Z. Hua, G. Lebanon, Y. Ma, P. Shivaswamy, H.-P. Tseng, J. Yang, et L. Zhang (2015). Personalizing linkedin feed. In *Proc. of the 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 1651–1660. ACM.

Alonso, O., C. C. Marshall, et M. Najork (2013). Are some tweets more interesting than others ?# hardquestion. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, pp. 2. ACM.

Bender, E. A. (2012). *An introduction to mathematical modeling*. Courier Corporation.

Berkovsky, S. et J. Freyne (2015). Personalised Network Activity Feeds : Finding Needles in the Haystacks. In *Mining, Modeling, and Recommending'Things' in Social Media*, pp. 21–34. Springer.

Berkovsky, S., J. Freyne, et G. Smith (2012). Personalized network updates : increasing social interactions and contributions in social networks. In *User Modeling, Adaptation, and Personalization*, pp. 1–13. Springer.

Bishop, C. M. (2006). Pattern recognition. *Machine Learning 128*.

Bontcheva, K., G. Gorrell, et B. Wessels (2013). Social media and information overload : Survey results. *arXiv preprint arXiv :1306.0813*.

Bourke, S., M. O'Mahony, R. Rafter, et B. Smyth (2013). Ranking in information streams. In *Proceedings of the companion publication of the 2013 international conference on Intelligent user interfaces companion*, pp. 99–100. ACM.

Chen, K., T. Chen, G. Zheng, O. Jin, E. Yao, et Y. Yu (2012). Collaborative personalized tweet recommendation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 661–670. ACM.

Cortes, C. et V. Vapnik (1995). Support-vector networks. *Machine learning 20*(3), 273–297.

Eckerson, W. W. (2007). Predictive Analytics. *Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report. Q 1*, 2007.

Ester, M. (2013). Recommendation in social networks. In *RecSys*, pp. 491–492.

Feng, W. et J. Wang (2013). Retweet or not ? : personalized tweet re-ranking. In *Proc. of the sixth ACM int. conf. on Web search and data mining*, pp. 577–586.

Freund, Y. et R. E. Schapire (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pp. 23–37. Springer.

Freyne, J., S. Berkovsky, E. M. Daly, et W. Geyer (2010). Social networking feeds : recommending items of interest. In *Proceedings of the fourth ACM conference on Recommender systems*, pp. 277–280. ACM.

Gilbert, E. et K. Karahalios (2009). Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 211–220. ACM.

Guy, I., I. Ronen, et A. Raviv (2011). Personalized activity streams : sifting through the river of news. In *Proceedings of the fifth ACM conference on Recommender systems*, pp. 181–188. ACM.

Hannachi, L., N. Benblidia, O. Boussaid, et F. Bentayeb (2015). Community Cube : a semantic framework for analysing social network data. *Int. Jour. of Metadata, Semantics and Ontologies 10*(3), 155–169.

Hong, L., R. Bekkerman, J. Adler, et B. D. Davison (2012). Learning to rank social update streams. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 651–660. ACM.

Krishnan, K. (2013). *Data warehousing in the age of big data*. Newnes.

Kuang, L., X. Tang, M. Yu, Y. Huang, et K. Guo (2016). A comprehensive ranking model for tweets big data in online social network. *EURASIP Journal on Wireless Communications and Networking 2016*(1), 1.

Lakkaraju, H., A. Rai, et S. Merugu (2011). Smart news feeds for social networks using scalable joint latent factor models. In *Proceedings of the 20th international conference companion on World wide web*, pp. 73–74. ACM.

Lomotey, R. K. et R. Deters (2014). Towards Knowledge Discovery in Big Data. In *Service Oriented System Engineering (SOSE), 2014 IEEE 8th Int. Symposium*, pp. 181–191.

McCulloch, W. S. et W. Pitts (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics 5*(4), 115–133.

Mohri, M., A. Rostamizadeh, et A. Talwalkar (2012). *Foundations of machine learning*. MIT-press.

Nyce, C. et A. CPCU (2007). Predictive analytics white paper. *American Institute for CPCU. Insurance Institute of America*, 9–10.

Oukid, L., O. Boussaid, N. Benblidia, et F. Bentayeb (2016). TLabel : A New OLAP Aggregation Operator in Text Cubes. *Int. Jour. of Data Warehousing and Mining 12*(4), 54–74.

Paek, T., M. Gamon, S. Counts, D. M. Chickering, et A. Dhesi (2010). Predicting the Importance of Newsfeed Posts and Social Network Friends. In *AAAI*, Volume 10, pp. 1419–1424.

Pan, Y., F. Cong, K. Chen, et Y. Yu (2013). Diffusion-aware personalized social update recommendation. In *Proc. of the 7th ACM conf. on Recommender systems*, pp. 69–76. ACM.

Rader, E. et R. Gray (2015). Understanding user beliefs about algorithmic curation in the

Facebook news feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 173–182. ACM.

Ramage, D., S. T. Dumais, et D. J. Liebling (2010). Characterizing Microblogs with Topic Models. *ICWSM 10*, 1–1.

Schwenk, H. et Y. Bengio (2000). Boosting neural networks. *Neural Computation 12*(8), 1869–1887.

Shen, K., J. Wu, Y. Zhang, Y. Han, X. Yang, L. Song, et X. Gu (2013). Reorder user's tweets. *ACM Transactions on Intelligent Systems and Technology (TIST) 4*(1), 6.

Soh, P.-H., Y.-C. Lin, et M.-S. Chen (2013). Recommendation for online social feeds by exploiting user response behavior. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 197–198. ACM.

Uysal, I. et W. B. Croft (2011). User oriented tweet ranking : a filtering approach to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 2261–2264. ACM.

Xu, Z., Y. Liu, N. Yen, L. Mei, X. Luo, X. Wei, et C. Hu (2016). Crowdsourcing based description of urban emergency events using social media big data. *IEEE Transactions on Cloud Computing*.

Yan, R., M. Lapata, et X. Li (2012). Tweet recommendation with graph co-ranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1*, pp. 516–525. Association for Computational Linguistics.

Zhan, L., Y. Sun, N. Wang, et X. Zhang (2016). Understanding the influence of social media on people's life satisfaction through two competing explanatory mechanisms. *Aslib Journal of Information Management 68*(3), 347–361.

Zheng, Z., K. Chen, G. Sun, et H. Zha (2007). A regression framework for learning ranking functions using relative relevance judgments. In *Proceedings of the 30th ann. inter. ACM SIGIR conf. on Research and development in information retrieval*, pp. 287–294.

Summary

Due to the large amount of information (messages, articles, videos, music, images, etc.) generated and shared on social networking sites, users find themselves overwhelmed by information generated chronologically in their news feed. In addition, most of information may be irrelevant. Sorting social updates, in order of relevance, is proposed as a solution to help users quickly view and interact with information that may interest them. In this work, we study existing approaches in the area of sorting social updates and expose their limits and some open issues according to several axes: factors influencing information's relevance, prediction models of information's relevance, training and evaluation of prediction models, etc.