Soutenance de Thèse

Institut de Recherche
en Informatique de Toulouse

Lundi 17 Septembre 2018
10h00
UT3 Paul Sabatier, IRIT, Auditorium J. Herbrand

Abdelhamid CHELLAL
Equipe IRIS - IRIT

Synthèse d'événement dans les médias sociaux : résumé rétrospectif et prospectif de microblogs

Jury :
- Lynda TAMINE-LECHANI – *Professeur, Université Toulouse 3 - Examinatrice*
- Sihem AMER-YAHIA – *Directrice de Recherche CNRS - Rapportrice*
- Patrice BELLOT – *Professeur, Université Aix-Marseille - Rapporteur*
- Philippe MULHEM – *Chargé de Recherche CNRS - Examinateur*
- Bernard DOUSSET – *Professeur émérite, Université Toulouse 3 - Directeur*
- Mohand BOUGHANEM – *Professeur, Université Toulouse 3 - Co-directeur*

Mots-clés : Recherche d'Information,Filtrage temps réel de tweets,Synthèse de tweets,Signes sociaux,Apprentissage adaptatif,Optimisation linéaire

Résumé : Le contenu généré dans les médias sociaux comme Twitter permet aux utilisateurs d'avoir un aperçu rétrospectif d'évènement et de suivre les nouveaux développements dès qu'ils se produisent. Cependant, bien que Twitter soit une source d'information importante, il est caractérisé par le volume et la vélocité des informations publiées qui rendent difficile le suivi de l'évolution des évènements. Pour permettre de mieux tirer profit de ce nouveau vecteur d'information, deux tâches complémentaires de recherche d'information dans les médias sociaux ont été introduites : la génération de résumé rétrospectif qui vise à sélectionner les tweets pertinents et non redondant récapitulant "ce qui s'est passé" et l'envoi des notifications prospectives dès qu'une nouvelle information pertinente est détectée. Notre travail s'inscrit dans ce cadre. L'objectif de cette thèse est de faciliter le suivi d'événement, en fournissant des outils de génération de synthèse adaptés à ce vecteur d'information. Les défis majeurs sous-jacents à notre problématique découlent d'une part du volume, de la vélocité et de la variété des contenus publiés et, d'autre part, de la qualité des tweets qui peut varier d'une manière considérable. La tâche principale dans la notification prospective est l'identification en temps réel des tweets pertinents et non redondants. Le système peut choisir de retourner les nouveaux tweets dès leurs détections où bien de différer leur envoi afin de s'assurer de leur qualité. Dans ce contexte, nos contributions se situent à ces différents niveaux : Premièrement, nous introduisons Word Similarity Extended Boolean Model (WSEBM), un modèle d'estimation de la pertinence qui exploite la similarité entre les termes basée sur le word embedding et qui n'utilise pas les statistiques de flux. L'intuition sous-jacente à notre proposition est que la mesure de similarité à base de word embedding est capable de considérer des mots différents ayant la même sémantique ce qui permet de compenser le non-appariement des termes lors du calcul de la pertinence. Deuxièmement, l'estimation de nouveauté d'un tweet entrant est basée sur la comparaison de ses termes avec les termes des tweets déjà envoyés au lieu d'utiliser la comparaison tweet à tweet. Cette méthode offre un meilleur passage à l'échelle et permet de réduire le temps d'exécution. Troisièmement, pour contourner le problème du seuillage de pertinence, nous utilisons

05 61 55 65 10
info@irit.fr
www.irit.fr

un classificateur binaire qui prédit la pertinence. L'approche proposée est basée sur l'apprentissage supervisé adaptatif dans laquelle les signes sociaux sont combinés avec les autres facteurs de pertinence dépendants de la requête. De plus, le retour des jugements de pertinence est exploité pour re-entrainer le modèle de classification. Enfin, nous montrons que l'approche proposée, qui envoie les notifications en temps réel, permet d'obtenir des performances prometteuses en termes de qualité (pertinence et nouveauté) avec une faible latence alors que les approches de l'état de l'art tendent à favoriser la qualité au détriment de la latence. Cette thèse explore également une nouvelle approche de génération du résumé rétrospectif qui suit un paradigme différent de la majorité des méthodes de l'état de l'art. Nous proposons de modéliser le processus de génération de synthèse sous forme d'un problème d'optimisation linéaire qui prend en compte la diversité temporelle des tweets. Les tweets sont filtrés et regroupés d'une manière incrémentale en deux partitions basées respectivement sur la similarité du contenu et le temps de publication. Nous formulons la génération du résumé comme étant un problème linéaire entier dans lequel les variables inconnues sont binaires, la fonction objective est à maximiser et les contraintes assurent qu'au maximum un tweet par cluster est sélectionné dans la limite de la longueur du résumé fixée préalablement.

05 61 55 65 10
info@irit.fr
www.irit.fr