Nouvelle approche de contextualisation de tweets basée sur les règles d'association inter-termes

Meriem Amina Zingla*, Mohamed Ettaleb**
Chiraz Latiri** Yahia Slimani*

*Université de Carthage, INSAT, laboratoire de recherche LISI , Tunis, Tunisie
** Université de Tunis El Manar, Faculté des Sciences de Tunis,
Tunis ,Tunisie

1 Introduction

Les tweets sont des messages courts ne dépassant pas 140 caractères. Cette contrainte impose l'utilisation d'un vocabulaire particulier pour les rédiger et donc elle rend indispensable de connaitre leurs contextes pour les comprendre. Pour ces raisons, nous allons nous concentrer sur la tâche de contextualisation des tweets attribuée à INEX2014 [1]. Les participants devaient fournir un contexte, pour permettre aux lecteurs de bien comprendre le tweet en utilisant un système de recherche d'information SRI et système de résumé automatique SRA. Dans cet article, nous proposons une nouvelle approche de contextualisation de tweets basée sur les règles d'association inter-termes.

Cet article est organisé comme suit : Dans la section 2, nous détaillons notre nouvelle approche. la section 3 sera consacrée aux différentes expériences menées , finalement nous conclurons dans la section 4.

2 Approche proposée

Les étapes suivantes décrivent le processus de contextualisation des tweets :

1. Sélection d'une sous collection d'articles à partir de la collection de documents.
2. Annotation des articles sélectionnés en utilisant TreeTagger [2].
3. Extraction des noms à partir des articles Wikipedia annotés.
4. Génération des règles d'association en utilisant l'algorithme CHARM [3] Zaki et Hsiao (2002).
5. Obtention de l'espace thématique de chaque tweet en projetant les tweets sur l'ensemble des règles d'association ;
6. Création de la requête à partir des mots de tweet et l'espace thématique.
7. L'envoi de la requête au système Baseline pour extraire le contexte de tweet.

1. https ://inex.mmci.uni-saarland.de/
2. http ://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/
3. http ://www.cs.rpi.edu/ zaki/www- new/pmwiki.php/Software/Software

3 Expérimentations et résultats

Nous présentons dans cette partie les résultats obtenus par notre système selon deux métriques : l'informativité de contexte (*voir Table* 1) et la lisibilité de contexte Bellot et al. (2013) (*voir Table* 2).

Run ID	Évaluation basée sur les sentences				Évaluation basée sur les phrases			
Run Id	Rang	Unigram	Bigram	Skip	Rang	Unigram	Bigram	Skip
361	**3**	**0.7632**	**0.8689**	**0.8702**	**3**	**0.7903**	**0.9273**	**0.9461**
360	**4**	**0.782**	**0.8925**	**0.8934**	**6**	**0.8104**	**0.9406**	**0.9553**
359	**7**	**0.8022**	**0.912**	**0.9127**	**8**	**0.8227**	**0.9487**	**0.9613**

TAB. 1 – *Évaluation sur le contenu informatif du contexte :INEX2014*

Run Id	Rang	lisible	Syntaxe	Diversité	Structure	Avg
360	**5**	**92.6%**	**70.35%**	**58.84%**	**86.33%**	**77.03%**
359	**8**	**93.03%**	**70.64%**	**53.53%**	**86.34%**	**75.88%**
361	**11**	**93.23%**	**70.41%**	**50.12%**	**85.97%**	**74.93%**

TAB. 2 – *Évaluation sur la facilité de lecture du contexte :INEX2014*

4 Conclusion

Dans cet article, nous avons décrit une nouvelle approche de contextualisation de tweet basée sur règles d'association inter-termes. Les résultats ont confirmé que la synergie entre les règles d'association entre termes et l'expansion de tweets est fructueuse. Dans un travail en cours, nous proposons d'ajouter une phase de désambiguïsation pour réduire le bruit dans nos résultats.

Références

Bellot, P., V. Moriceau, J. Mothe, E. SanJuan, et X. Tannier (2013). Overview of INEX tweet contextualization 2013 track. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*

Zaki, M. et C.-J. Hsiao (2002). An efficient algorithm for closed itemset mining. In *Second SIAM International Conference on Data Mining*.

Summary

Tweets are short messages that do not exceed 140 characters. Since they must be written respecting this limitation, a particular vocabulary is used. To make them understandable to a reader, it is therefore necessary to know their context. In this paper, we describe our approach for the tweet contextualization. This approach allows the extension of the tweet's vocabulary by a set of thematically related words using mining association rules between terms.