

Event Detection and Tracking

Proposta de Dissertação de Mestrado

Orientador: Sebastião Pais

Co-Orientador: João Paulo Cordeiro

Departamento de Informática @ UBI

2019/2020

1 Context

Social media have emerged as powerful means of communication for people looking to share and exchange information on a wide variety of real-world events. Short messages posted on Twitter can typically reflect these events as they happen. For this reason, the content of such social media sites is particularly useful for real-time identification of real-world events and their associated user-contributed messages. As such, a crowd can be viewed as a community sharing a common focus about some specific event.

Event detection has intensively been studied in the last decade mainly due to the advent of social media Farzindar et Inkpen (2017). Two different approaches have been proposed: document-pivot and feature-pivot. All the studied techniques are interesting but do not cover the overall picture. Once an event is detected, it is crucial to track it, i.e. to follow its evolution. Within this scope, topic models have shown successful results for text clustering tasks. However, they only rely on a term-document matrix to compute similarity, which may be insufficient for social media texts that are short and lack in contextual information. This is confirmed by the recent work of Wold et Vikre (2015), who show that the use of locality-sensitive hashing combined with named entity recognition achieves better performance for detecting news than using the topic modeling approach. Moreover, topics are represented as sets of words, that may not all be bursty, and thus may include more general topics than specific ones.

2 Goals

This thesis propose a new strategy based on the recent findings of Moreno *et al.* (2014) which proposed the Dual C-means clustering algorithm allowing to mix document-pivot and feature-pivot techniques into a unique model. Dual C-means showed to perform likewise topic models for word sense induction Acharya *et al.* (2016). The advantage of the Dual C-means algorithm is that different similarity measures can be implemented, including knowledge-based metrics, that may lead to improved results in the line of Wold et Vikre (2015), as tweets are linked to YAGO by their entities. Moreover, the clustering process can be driven by bursty

keywords or named entities, and thus better adapt to the dynamic environment of Twitter, instead of relying on all possible terms present in the time window. It will also allow the integration of richer sources of knowledge in clustering. One possibility is to explore entailment dependencies according to recent work from the Pais *et al.* (2011).

3 Tasks

1. Review the State-of-the-art, write the survey about this problematic;
2. Propose a new Unsupervised and Language Independent Approach to Event Detection and Tracking;
3. Implementation;
4. Testing and evaluation;
5. The writing of the dissertation.

4 Contactos

Sebastião Pais (sebastiao@di.ubi.pt) - Gabinete 4.1

João Paulo Cordeiro (jpaulo@di.ubi.pt) - Gabinete 4.3

UBI, Departamento de Informática
Rua Marquês d'Ávila e Bolama
6201-001 Covilhã

Bibliography

- ACHARYA, S., EKBAL, A., MORENO, J. G., SAHA, S., DIAS, G. et SANTHANAM, P. (2016). Multi-Objective Optimization for Word Sense Induction based on Content and Interlink Connections. *In 21st International Conference on Applications of Natural Language to Information Systems (NLDB 2016)*, Salford, United Kingdom.
- FARZINDAR, A. et INKPEN, D. (2017). Natural language processing for social media, second edition. *Synthesis Lectures on Human Language Technologies*, 10(2):1–195.
- MORENO, J. G., DIAS, G. et CLEUZIQU, G. (2014). Query log driven web search results clustering. *In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 777–786, New York, NY, USA. ACM.
- PAIS, S., DIAS, G., WEGRZYN-WOLSKA, K., MAHL, R. et JOUVELOT, P. (2011). Textual entailment by generality. *Procedia - Social and Behavioral Sciences*, 27:258 – 266. Computational Linguistics and Related Fields.
- WOLD, H. M. et VIKRE, L. C. (2015). Online news detection on twitter. Mémoire de D.E.A., NTNU.