ARC'18
مؤتمر مؤسسة قطر السنوي للبحوث
QATAR FOUNDATION ANNUAL RESEARCH CONFERENCE
البحث والتطوير: التركيز على الأولويات، وإحداث الأثر
R&D: FOCUSING ON PRIORITIES, DELIVERING IMPACT
19-20 مـــارس
19-20 MARCH

مـؤسـسـة قـطـر
Qatar Foundation
لإطـلاق قـدرات الإنـسـان.
Unlocking human potential.

Computing & Information Technology - Poster Display

http://doi.org/10.5339/qfarc.2018.ICTPD1061

Effective Realtime Tweet Summarization

Reem Suwaileh*, Tamer Elsayed

Qatar University, College of Computer Science and Engineering.
* reem.suwaileh@qu.edu.qa

Twitter has been developed as an immense information creation and sharing network through which users post information. Information could vary from the world›s breaking news to other topics such as sports, science, religion, and even personal daily updates. Although a user would regularly check her Twitter timeline to stay up-to-date on her topics of interest, it is impossible to cope with manual tracking of those topics while tackling the challenges that emerge from the Twitter timeline nature. Among these challenges are the big volume of posted tweets (about 500M tweets are posted daily), noise (e.g., spam), redundant information (e.g., tweets of similar content), and the rapid development of topics over time. This necessitates the development of real-time summarization systems (RTS) that automatically track a set of predefined interest profiles (representing the users› topics of interest) and summarize the stream while considering the relevance, novelty, and freshness of the selected tweets. For instance, if a user is interested in following the updates on the "GCC crises», the system should efficiently monitor the stream and capture the on-topic tweets including all aspects of the topic (e.g., official statements, interviews and new claims against Qatar) which change over time. Accordingly, real-time summarization approaches should use simple and efficient approaches that can scale to follow multiple interest profiles simultaneously. In this work, we tackle such problem by proposing RTS system that adopts a lightweight and conservative filtering strategy. Given a set of user interest profiles, the system tracks those profiles over Twitter continuous live stream in a scalable manner in a pipeline of multiple phases including pre-qualification, preprocessing, indexing, relevance filtering, novelty filtering, and tweets nomination. In prequalification phase, the system filters out non-English and low-quality tweets (i.e., tweets that are too short or including many hashtags). Once a tweet is qualified, the system preprocesses it in a series of steps

دار جامعة حمد بن خليفة للنشر
HAMAD BIN KHALIFA UNIVERSITY PRESS

Cite this article as: Suwaileh R and Elsayed T. (2018). Effective Realtime Tweet Summarization. Qatar Foundation Annual Research Conference Proceedings 2018: ICTPD1061 http://doi.org/10.5339/qfarc.2018.ICTPD1061.

(e.g., removing special characters) that aim at preparing the tweet for relevance and novelty filters. The system adopts a vector space model where both interest profiles and incoming tweets are represented as vectors constructed using idf-based term weighting. An incoming tweet is scored for relevance against the interest profiles using the standard cosine similarity. If the relevance score of a tweet exceeds a predefined threshold, the system adds the tweet to the potentially-relevant tweets for the corresponding profile. The system then measures the novelty of the potentially-relevant tweet by computing its lexical overlap with the already-pushed tweets using a modified version of Jaccard similarity. A tweet is considered novel if the overlap does not exceed a predefined threshold. This way the system does not overwhelm the user with redundant notifications. Finally, the list of potentially-relevant and novel tweets of each profile is re-ranked periodically based on both relevance and freshness and the top tweet is then pushed to the user; that ensures the user will not be overwhelmed with excessive notifications while getting fresh updates. The system also allows the expansion of the profiles over time (by automatically adding potentially-relevant terms) and the dynamic change of the thresholds to adapt to the change in the topics over time. We conducted extensive experiments over multiple standard test collections that are specifically developed to evaluate RTS systems. Our live experiments on tracking more than 50 topics over a large stream of tweets lasted for 10 days show both effectiveness and scalability of our system. Indeed, our system exhibited the best performance among 19 international research teams from all over the world in a research track organized by NIST institute (in the United States) last year.