論 文 の 内 容 の 要 旨

論文題目　　Identifying and Summarizing Public Opinion to Real-time Events Using Microblogs as Social Sensors
(ソーシャルセンサとしてのマイクロブログを利用したリアルタイムイベントへの世論の識別と要約)

氏　　名　　モハンマド　アシフ　ホッセン　カン

Twitter has become an ideal platform for getting access to first-hand responses from its users in real-time about various social events, for example, popular public events such as Super Bowl, Academy Awards, presidential debates etc., social movements such Arab Spring, natural disasters such as earthquakes, tsunamis etc., outburst of epidemics such as seasonal flu and so on. With its 200 million monthly active users posting over 500 million tweets per day, Twitter offers a means to harness collective distributed opinion of the crowd regarding various social issues. It is as though a huge network of social sensors have been deployed, where individual sensors gets activated and communicates with other sensors whenever it perceives some social signals from the segment of the society where it is embedded. However, the sheer volume of these real-time status updates presents a great challenge in identifying, filtering, distilling and summarizing tweets pertinent to any particular event. Moreover, the 140 character length restriction imposed by Twitter on individual posts, forces its users to use unstructured language, non-grammatical sentences, abbreviations and out of dictionary words. This makes the aforementioned challenges even more complicated as many of the tools and techniques developed for standard text processing in the field of Natural Language Processing, cannot be readily imported in the domain of Twitter.

In the first part of this dissertation we propose a method for identifying tweets relevant to any particular event using *Bag-of-Word* classifiers. We investigate a special scenario where the training set is not carefully annotated (e.g. when “hashtags” are used as class labels). Although, it alleviates the need for manual labeling of training instances, thus saving time and cost, it presents several challenges. One of the challenges we tackled is the identification and removal of “confounding outliers”, i.e. training instances from a class which are outliers in their own classes and are also very similar to instances from other classes. Evaluation shows that their removal together with our proposed feature selection method can improve the classification accuracy. The proposed method can also improve classification accuracy when the training set

contains limited training instances and the inter-class distances are very narrow.

In the second part of the thesis, we propose an unsupervised method for automatically summarizing tweets related to a particular event. The proposed method can produce a journalistic summary for those events which are pre-scheduled, take place within a limited time span, possibly televised and cause upsurge of traffic in Twitter. The proposed method tries to optimize three objective functions: maximize information content, minimize redundancy and maximize topical diversity. It optimizes the objective functions independently; i.e. in succession. At first, it performs topical clustering of the tweets; i.e. cluster the tweets according to their discussion topics. It performs a hard clustering of the tweets in the collection. Then, salience scores of the tweets in each topic cluster are calculated independently. Next, a set of highly salient tweets from each cluster are incorporated in the summary for that cluster. However, to avoid redundancy, tweets similar to the already selected tweets in the summary are not included in the summary. Finally, tweets in the summary of each cluster are aggregated to produce the summary of the whole corpus. Evaluation performed on thousands of relevant tweets generated during a real-world event reveals that, the summary produced by the proposed model could delineate the proceeding of the event with high precision and recall and could significantly outperform two intuitive and competitive baseline methods.

In the third part of the thesis, we propose another unsupervised method for automatically summarizing even-related tweets. However, instead of optimizing the objective functions independently, we model the optimization problems using Integer Linear Programming (ILP) and jointly optimize them using Lagrangian relaxation and dual decomposition techniques. Instead of trying to perform a hard clustering to tweets in the collection, we try to identify the inclination of each tweet toward the prominent discussion topics. As there are no topic clusters now, we calculate salience scores of the tweets in the whole collection instead of doing it per cluster basis. Two intermediate summaries are produced for the tweet collection: one trying to maximize topical diversity and the other trying to maximize salience score. However, the one trying to maximize salience score performs sequential selection of tweets in the summary and uses maximum marginal relevance (MMR) to penalize the salience score of a tweet in proportion to the already incorporated tweets in the summary. Finally, Lagrangian heuristics are used to find an agreement between the two optimizers to produce the final summary. Evaluation performed on the same data set used in the second part shows that the joint optimization approach significantly outperforms the sequential optimization approach that we proposed earlier.

Thus, the thesis as a whole presents a framework for identifying and summarizing people's opinion expressed in their tweets about public events. Twitter users post over two billion queries each day looking for real-time status updates on different events. However, the sheer volume of

tweets pertinent to a particular event is formidable for a search user to go though even a fraction of them. On November 2013, Twitter introduced a new tool called the “custom timelines” that allows its users to manually select tweets returned by Twitter's search interface in response to an event related query to construct custom lists of tweets on their topic of interest. This substantiates that the proposed framework for automatically generating summary of an event by identifying tweets with popular discussion points among the set of pertinent tweets is highly significant and would satisfy the information need of a large user group.