Temporal Classification and Visualization of Topics in a Twitter Search Interface

Radhika Gopi*
University of Regina

Orland Hoeber†
University of Regina

Abstract

Searching within Twitter is a challenging task; the short and cryptic nature of tweets leads to search results sets that may include information on many different topics. While many topic modelling approaches exist to extract the salient topics from the tweets, what is missing is a method for temporally classifying the topics and showing these to a searcher to help them understand the makeup of the search results. In this work, we model the temporal distribution of the tweets that match each extracted topic, and then classify the topic as either emergent, stable, waning, or unclassified. This classification is visualized along side each topic, giving the searcher an easy way of identifying how the topics are changing over time. An example of this approach within an existing mobile Twitter search interface is provided.

Index Terms: H.3.2 [Information Storage and Retrieval]: Information Search and Retrieval—Information filtering H.5.2 [Information Interfaces and Presentation]: User Interfaces—Screen design;

1 Introduction

User generated content services such as Twitter make it easy for people to post and browse messages on the go. They serve as a crucial source of public opinion on a wide range of topics. However, the ability to search for information within Twitter is limited by the query box and search results list paradigm. The user has to read through the entire chronologically ordered list of tweets in order to understand the potentially broad range of information shared for a particular search query. The list based representation can be effective when presenting search results that are narrowly focused on a specific topic, but less effective when there is some degree of ambiguity among the information displayed.

Despite the limitations of mobile devices (e.g., limited memory, processing power, and screen size), people are still interested to utilize mobile devices for complex information seeking tasks. When designing a mobile search app, it is beneficial to make use of visual and interactive techniques to allow the searcher to focus their efforts on their intended topic. In prior work, we developed a mobile search app called TwIST(Twitter Information Search Tool) for searching and exploring within Twitter [4]. This tool supports search and exploration among tweets matching user-specified queries, using topic modelling and topic visualization to support exploratory search tasks. TwIST was designed to allow searchers to easily identify and navigate among the information presented, and to focus their efforts on evaluating topics of interest.

Topics are automatically extracted from the set of tweets in the search results, and are summarized with a label and a visual indicator of how many tweets are on that topic. The tweets are supplemented with the titles of any embedded URLs, and are clustered a hybrid approach that consists of k-nearest neighbours (kNN), part of speech tagging, n-gram extraction, and text filtering. TwIST allows a searcher to interactively tap on the topics of interest in order to filter the search results. When multiple topics are selected, a compact visualization of the tweet-topic correspondence is provided to the right of the list to compare the topics to one another.

Given the importance of the temporal nature of the tweets, TwIST ranks the tweets by their creation time in descending order and displays the latest tweet on the top. This order is helpful to searchers since they generally wish to know what people are tweeting about at the current point in time. The primary contribution of this paper is to enhance the topic modelling approach used in TwIST with a mechanism for classifying the topics based on the temporal distribution of the tweets in each topic. By providing a visual indication of this classification, searchers will be able to readily identify those topics that are *emergent*, *stable*, *waning*, from those that have no identifiable temporal pattern.

2 Related Work

Topic extraction from user generated content is an interesting knowledge discovery problem. Among the various topic models that have been discussed in the literature, Latent Dirichlet Allocation (LDA) [2] is a widely used model. Unfortunately, for short and cryptic textual content such as that posted on Twitter, there is often not enough content available upon which to base the topic model. While some have augmented Twitter data with information from external resources [1], doing so comes at the cost of slowing down the topic modelling process. For a search interface, it is important for this process to be able to complete in near-real time. To this end, we have developed an approach that uses fast kNN clustering and POS tagging, and offloading the work to a dedicated server [4].

A number of research projects have been initiated to analyze the temporal distribution of tweets [5, 6]. Of interest is the work by Lee. et al., who classified trending topics as long-term, mid-term, or short-term, based on the temporal signature of the tweets [3]. Unfortunately, such approaches are often computationally intensive and make use of a large collection of tweets. For the purposes of search, both the amount of data available and the amount of time a searcher is willing to wait are constrained.

3 Temporal Classification of Topic Distributions

When a query is submitted within the TwIST app, the server retrieves the top 100 tweets from the Twitter API [7], and passes these to the mobile application as-is. It then runs a topic modelling algorithm to extract the topics out of the search result set. In order to support the classification of the temporal distribution of the tweets, the full temporal range of the search results set is calculated and is divided into n equal sized bins. For the current implementation, a value of $n = 4$ was determined to be sufficient for the relatively course classification of the temporal patterns (e.g., *emergent*, *stable*, *waning*, and *unclassified*).

For each topic, the total number of tweets were counted for each of the four temporal bins. This data was then normalized over all of the tweets that were present for each temporal bin, resulting in a percent of tweets in each bin for each topic. Using a cut-off of 5%, each topic was then classified based according to the following

*e-mail: gopi200r@uregina.ca

†e-mail: orland.hoeber@uregina.ca

rules (a) if the occurrence frequency of tweets is greater than cut-off in the most recent bins, but less in the older bins, then topic was classified as *emergent*; (b) if the occurrence frequency of tweets is greater than cut-off in the older bins, but less in the most recent bins, then the topic was classified as *waning*; (c) if the tweets occurrence was consistent among the bins, then the topic was classified as *stable*; (d) otherwise, the topic was *unclassified*.

4 Topic Visualization and Search Results Exploration

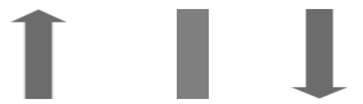
In order to allow the searcher to make use of the temporal classification of the topics, it is important to provide an easy to interpret visual indication of this information. The glyphs used for this purpose are shown in Figure 1, with an up-arrow indicating an emergent topic, a down arrow indicating a waning topic, a vertical bar indicating a stable topic, and an empty space indicating an unclassified topic. These graphical indicators are displayed beside each topic, and further enhance the existing TwIST indicator that shows the relative number of tweets for each topic.

The value of such an addition to TwIST is that it provides to the searcher further information about the topics that are embedded within the search results, making it easier for the searcher to make sense of what people are tweeting about and the temporal pattern of the topics. Since colour encoding is already used extensive to visually separate multiple topics that the searcher may select, it is not used for these indicators to avoid visual clutter.

An example of the search interface during an exploratory search activity is show in Figure 2. Suppose a searcher is interested in what people are saying about their favourite sport, cricket. A good stating query for this topic is #cricket, which will include tweets about many different topics related to this sport. These topics are extracted, along with their temporal classification. One can see from the topic list and supplemental visual encoding that many people are consistently talking more about Pakistan, as the topics "Miandad regarding Shahid", "PEMRA issues advices", and "Pakistan women cricket" are shown with solid bars indicating that they are stable topics. However, it is also easy to identify that the topic "India kick start is emergent, and "Pakistan beat Srilanka" is waning. By providing for each topic a temporal classification indicator, the searcher can glance at the topics and easily identify the temporal distribution. With the previous version of TwIST, in order to identify this information, a searcher would had to explore each topic independently and consider the features of the timeline visualization (shown at the top of the tweet list). This simple addition provides additional information to the searcher about the makeup of the search results and the features of the topics, allowing them to make informed choices about the topics they wish to explore.

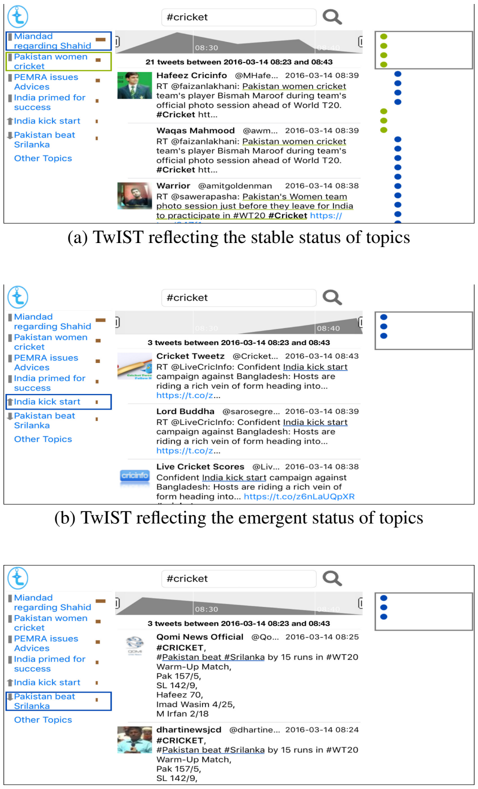
5 Conclusion

In this paper, we have presented our work on classifying topics based on the temporal distribution of tweets in Twitter search result set. By providing a visual indication of the temporal flow emergent, stable, waning, or unclassified, the searcher can readily use this knowledge when choosing which topics to explore. In future work, we will experiment with the number of temporal bins used to classify the tweets, and compare this heuristic-based approach to that of more complex temporal classification approaches. User evaluations of TwIST are currently in the planning stages.



(a) Emergent (b) Stable (c) Waning (d) Unclassified

Figure 1: Temporal classification indicators



#cricket

Miandad regarding Shahid
Pakistan women cricket
PEMRA issues Advices
India primed for success
India kick start
Pakistan beat Srilanka
Other Topics

21 tweets between 2016-03-14 08:23 and 08:43

Hafeez Cricinfo @MHafe... 2016-03-14 08:39
RT @faizanlakhani: Pakistan women cricket team's player Bismah Maroof during team's official photo session ahead of World T20. **#Cricket** htt...

Waqas Mahmood @awm... 2016-03-14 08:39
RT @faizanlakhani: Pakistan women cricket team's player Bismah Maroof during team's official photo session ahead of World T20. **#Cricket** htt...

Warrior @amitgoldenman 2016-03-14 08:38
RT @sawerapasha: Pakistan's Women team photo session just before they leave for India to practicipate in #WT20 **#Cricket** https://

(a) TwIST reflecting the stable status of topics

#cricket

Miandad regarding Shahid
Pakistan women cricket
PEMRA issues Advices
India primed for success
India kick start
Pakistan beat Srilanka
Other Topics

3 tweets between 2016-03-14 08:23 and 08:43

Cricket Tweetz @Cricket... 2016-03-14 08:43
RT @LiveCricInfo: Confident India kick start campaign against Bangladesh: Hosts are riding a rich vein of form heading into... https://t.co/z...

Lord Buddha @sarosegre... 2016-03-14 08:39
RT @LiveCricInfo: Confident India kick start campaign against Bangladesh: Hosts are riding a rich vein of form heading into... https://t.co/z...

Live Cricket Scores @Liv... 2016-03-14 08:38
Confident India kick start campaign against Bangladesh: Hosts are riding a rich vein of form heading into... https://t.co/z6nLaUQpXR

(b) TwIST reflecting the emergent status of topics

#cricket

Miandad regarding Shahid
Pakistan women cricket
PEMRA issues Advices
India primed for success
India kick start
Pakistan beat Srilanka
Other Topics

3 tweets between 2016-03-14 08:23 and 08:43

Qomi News Official @Qo... 2016-03-14 08:25
#CRICKET,
#Pakistan beat #Srilanka by 15 runs in #WT20
Warm-Up Match,
Pak 157/5,
SL 142/9,
Hafeez 70,
Imad Wasim 4/25,
M Irfan 2/18

dhartinewsjcd @dhartine... 2016-03-14 08:24
#CRICKET,
#Pakistan beat #Srilanka by 15 runs in #WT20
Warm-Up Match,
Pak 157/5,
SL 142/9,

(c) TwIST reflecting the waning status of topics

Figure 2: An example of using TwIST to isolate tweets about specific topics

References

[1] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. Eddi: Interactive topic-based browsing of social status streams. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*, pages 303–312, 2010.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] M. Cheong and V. Lee. Integrating web-based intelligence retrieval and decision-making from the Twitter trends knowledge base. In *Proceedings of the Workshop on Social Web Search and Mining*, pages 1–8, 2009.

[4] R. Gopi and O. Hoeber. TwIST: A mobile approach for searching and exploring within Twitter. In *Proceedings of the International Conference on Human Information Interaction and Retrieval*, pages 273–276, 2016.

[5] C.-H. Lee, H.-C. Yang, T.-F. Chien, and W.-S. Wen. A novel approach for event detection by mining spatio-temporal information on microblogs. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pages 254–259, 2011.

[6] R. D. Perera, S. Anand, K. Subbalakshmi, and R. Chandramouli. Twitter analytics: Architecture, tools and analysis. In *Proceedings of the Conference on Military Communications*, pages 2186–2191, 2010.

[7] Twitter. Twitter search API. `https://api.twitter.com/1.1/`. Accessed: 2016-03-14.