A Framework for Information Discovery from Twitter

Pooria Madani and Ali A. Ghorbani

Faculty of Computer Science, University of New Brunswick

Motivation

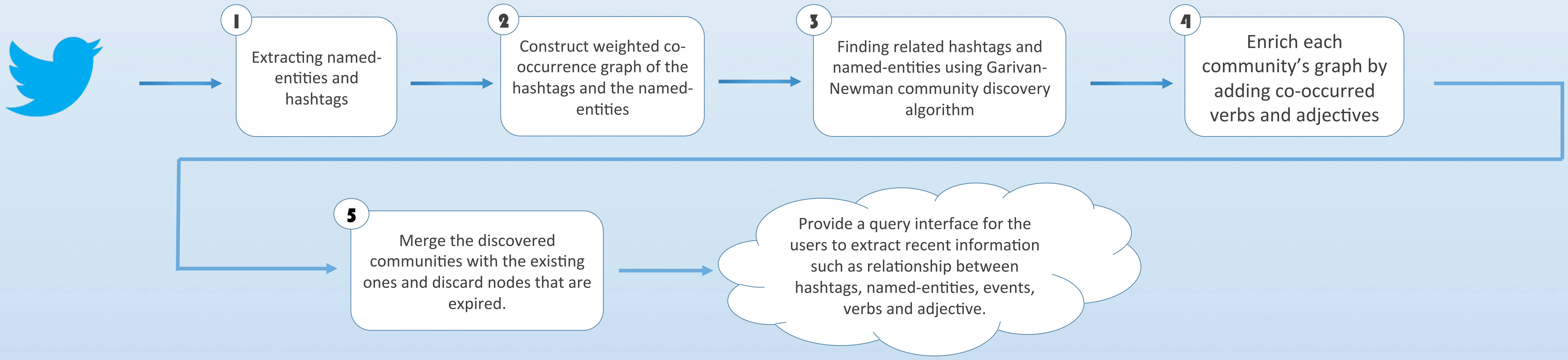
Users on Twitter have developed a tagging culture of placing a hash symbol (#) in front of short strings, called hashtags, on their posted messages, called tweets. Hashtags can represent different topics such as person, event, opinion, etc. Different hashtags may represent the same phenomenon. Thus, discovering relationship between hashtags may unveil further information about the trending topics in the Twitter and social media and can help researchers to perform more in depth analysis on tweets. Moreover, some hashtags represent opinion which can evolve through time. Discovering representative mood for a hashtag is only possible by discovering related opinion terms that co-occurred with the hashtag. Our purposed framework enables such discoveries and can help in analyzing posts in Twitter or any other social media source that utilizes hashtags in their system.

Solution

Building co-occurrence matrix of words for a set of documents for discovering relationship between the words, has proven to be a useful exercise in the field of text mining. However, due to the noise and size of available tweets, co-occurrence matrix of tweets' words needs to be constructed in multiple steps in order to correctly capture the clusters of trending phenomenon. Discovering relationships between hashtags and named-entities helps in resolving ambiguity in tasks such as tweet classification and clustering and increases their accuracy.

We have proposed a framework for constructing a graph that captures relationships between hashtags and other terms that are co-occurred within the tweets based on the frequency of occurrence. And providing API for researchers and developers to traverse the constructed graph for gaining further insight for tweet processing task.

Our Approach



1 Extracting named-entities and hashtags

2 Construct weighted co-occurrence graph of the hashtags and the named-entities

3 Finding related hashtags and named-entities using Garivan-Newman community discovery algorithm

4 Enrich each community's graph by adding co-occurred verbs and adjectives

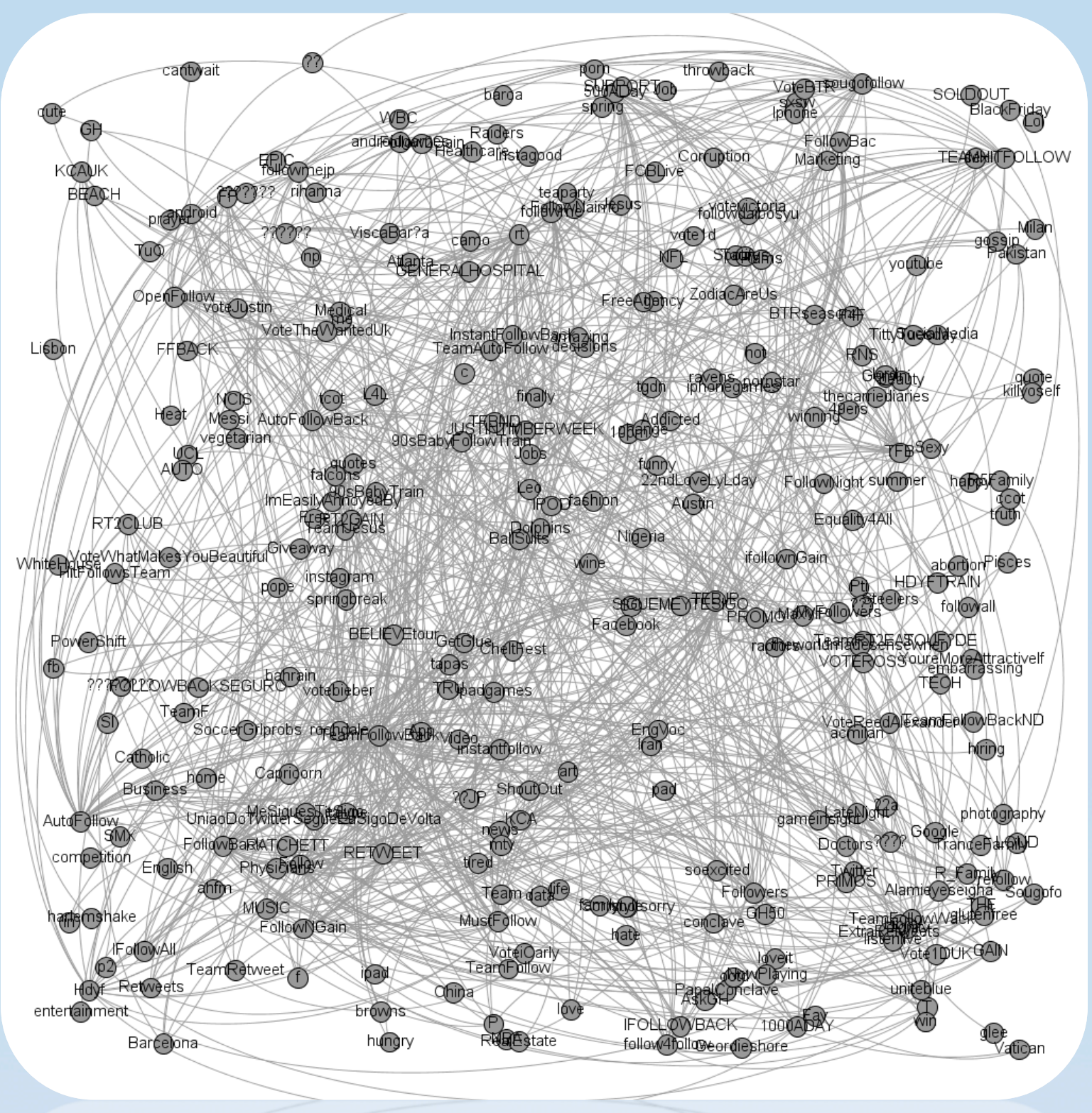
5 Merge the discovered communities with the existing ones and discard nodes that are expired.

Provide a query interface for the users to extract recent information such as relationship between hashtags, named-entities, events, verbs and adjective.

Hashtag and Named Entity Graph

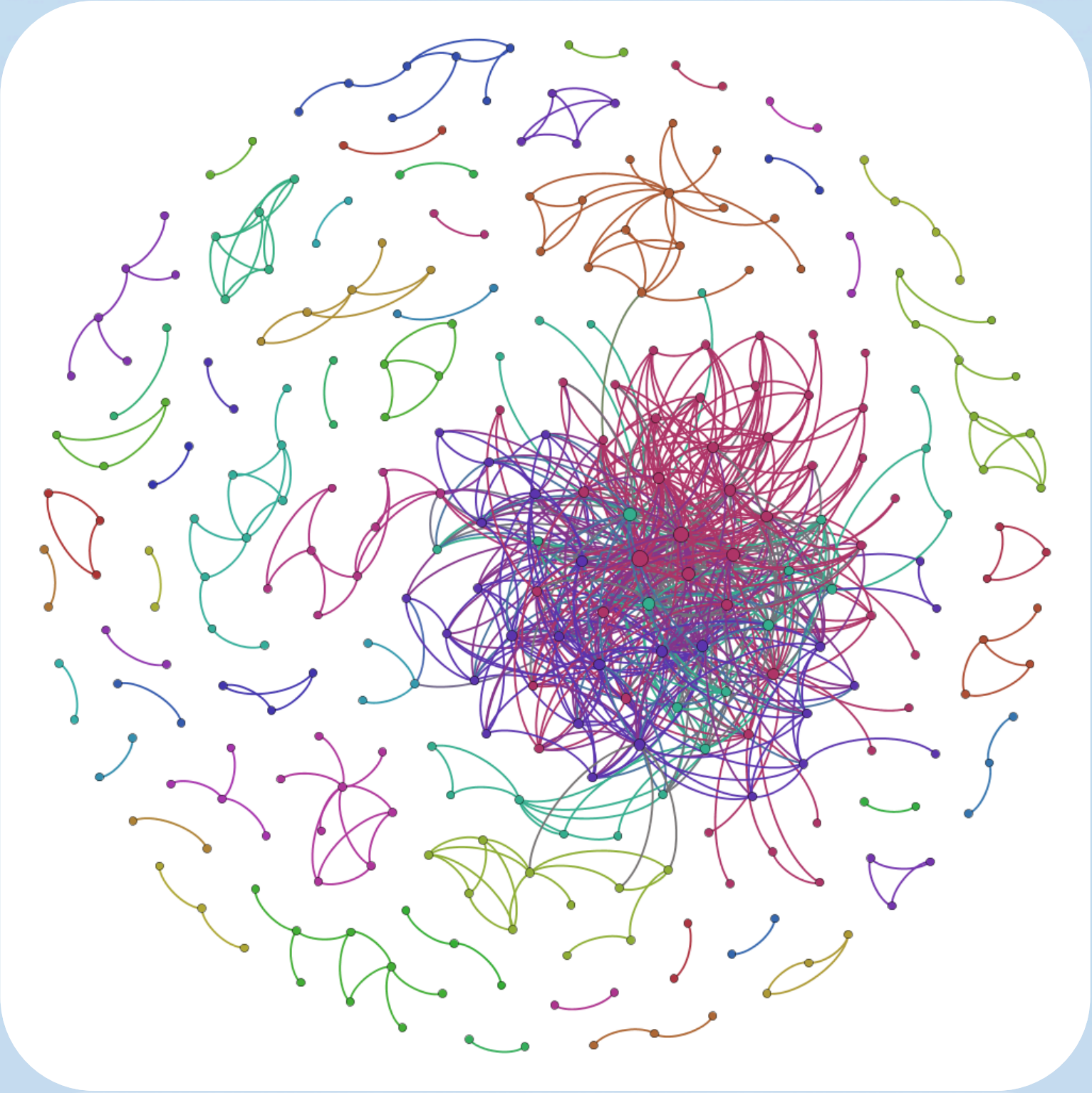
In order to discover meaningful hashtags communities, it is reasonable to enrich the hashtags co-occurrence graph with the named entities.

Using StandfordNLP software package we extract named-entities from tweet texts and add them to the graph. As shown in the figure, the result is a massively interconnected graph representing relationships between hashtags and named-entities with their frequency of co-occurrence recorded on the edges. However, due to the presence of noise, not all the discovered edges are meaningful. In order to boost the community discovery computation and its accuracy, we prune the graph's edges using minimum cut algorithm. Minimum cut algorithm eliminates edges that are less important in the graph based on the frequency and their connectedness.



Community Discovery

Girven-Newman algorithm is a method for detecting communities in complex systems. This algorithm defines communities by trying to find edges that are mostly located between communities; communities are detected by progressively deleting such edges. As shown in the figure, the result of this algorithms is set of hashtag-named-entity communities that are distinguished by colour. Once communities are detected, each community's graph gets further enrichment by adding co-occurred verbs, nouns, and adjectives. Each community's graph will be used to extract information such as events, locations, moods, and sentiment from the community by simply traversing the constructed graph.



2013 Faculty of Computer Science Research Expo. April 11-12, Wu Conference Centre, University of New Brunswick