Improving Hashtag Comprehension with Search and Text Summarization

John Lanchantin
University of Virginia
85 Engineer's Way
Charlottesville, VA
22904-4740
jjl5sw@virginia.edu

Nicholas Janus
University of Virginia
85 Engineer's Way
Charlottesville, VA
22904-4740
ncj2ey@virginia.edu

Weilin Xu
University of Virginia
85 Engineer's Way
Charlottesville, VA
22904-4740
xuweilin@virginia.edu

ABSTRACT

Posts on micro-blogging sites are often very hard to understand due to their informality. Hashtags represent one solution to this problem by acting as subject markers for posts. However, hashtags are often difficult to understand without reading through multiple posts or conversations. We attempt to solve hashtag comprehension problem by automatically understanding what hashtags mean, and displaying relevant documents or text from within those documents.

Categories and Subject Descriptors

Information systems [**Information Retrieval**]: Specialized information retrieval

Keywords

Micro-blogging, Hashtag retrieval, hashtag prediction, hashtag comprehension

1. INTRODUCTION

The Internet today, especially social network services such as twitter and facebook, is filled with 'hashtags'. Hashtags are single tokens that use the character '#' in front of the words, and are often composed of natural language n-grams or abbreviations. The problem is that there is no structure to hashtags beyond the format of the '#' character and no spaces, thus making it terribly difficult to understand them. Users often create hashtags that are slang, concatenations of many words, acronyms, or simply made up words.

An important task is to be able to automatically understand the underlying meaning behind a trending topic on social media. By understanding the meaning behind hashtags, we can further analyze what people are talking about. Often times, it does not become clear what someone is talking about until the meaning of his/her hashtag is understood. This leads to two important ideas: making it easier to quickly understand what someone is saying on social media, and also being able to do tasks such as document recommendation.

In order to understand the meaning behind hashtags, we propose two possible solutions. The first solution is extracting meaning from search engines by selecting proper keywords and searching on websites such as wikipedia, urbandictionary, Google News, etc. We will search several combinations of keywords (which are extracted from tweets that include the same hashtag) on existing search engine services like Google or Bing, gather the top results, and rank them to get the best document. We will then extract the key sentence/phrase from the most relevant document. The second solution is extracting meaning directly from tweets by using techniques such as graph mining, or automatic text summarization on a large amount of tweets with the same hashtag.

2. RELATED WORK

The research surrounding hashtags covers a variety of topics including hashtag retrieval[**?**], hashtag prediction[**?**][**?**] for social media posts, or sentiment analysis for either the hashtag itself or the contents of the enclosing post. Although these systems rely on an implicit understanding of the hash tag's meaning, they do not attempt to export such semantics. Attempts to deliver the meaning of hashtags rely on crowdsourcing or manual annotation.

Currently, twitter manually expands certain acronym hashtags (e.g. #oitnb translates to ?orange is the new black?), but it does not explain what they mean. In addition, there are a few websites (e.g. tagdef.com) which attempt to explain hashtags by crowdsourcing definitions, but there is often no clear cut definition. Both of the aforementioned methods do a poor job at explaining trending, or newly defined hashtags because they rely on the meaning of a hashtag over a long period of time, or have to wait until users explain the hashtag. In addition, the content quality is not as good as webpage links that give a more clearly defined explanation. Our approach will have a much greater coverage of hashtags and will not rely on manual intervention for an interpretation of the tag's meaning. By giving access to our system via a browser plugin, our service will also be much more accessible and available to users of micro-blogging platforms.

3. APPROACH AND IMPLEMENTATION

The main idea of our approach is as follows. Based on all tweets that use the inquired hashtag, we create a query for a current search engine (Bing.com), and display relevant web pages and a short summary based on those web pages to the user.
The system flow works in the following way: the user issues a hashtag to find its meaning using a browser plugin, the system extracts and filters all tweets using that hashtag, generates a query based on the tweets, searches that query using bing, returns the links of the most relevant web pages, and generates a summarization of the text from the top web page. The links and summary are displayed to the user via the plugin. This is shown in figure 1. Each module is explained is explained in further detail in the following subsections.

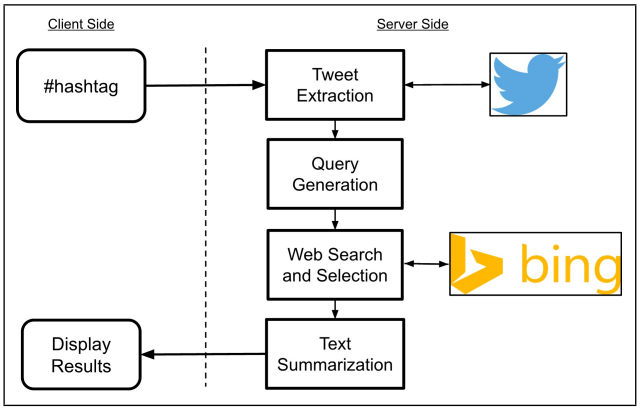

Figure 1: System Architechture

3.1 Tweet Extraction

Idea: return the text of N most recent tweets that use the inquired hashtag Limitation: Twitter Search API only allows tweets from past 6-9 days Filtering the tweets Removal of URLs, emoticons, spam tweets Case sensitive hashtags (e.g. #IAD vs #iAd) Return related hashtags

3.2 Query Generation

The main idea of the query generation module is to generate a query to search in Bing based on the text from the tweet extraction module. Our method generates a query based on the high-frequent terms from the tweets.
We first pre-process the filtered tweet text by using tokenization, case normalization, removing stopwords (which we added certain "twitter-specific" stopwords to the english stopwords list) and punctuation removal. We do not use stemming because... We also do not use segmentation because we found out that segmentation did not improve our results, and in some cases decreased accuracy. We assume that this is due to the fact that when we search something in Bing that contains concatenated words (e.g. 'presidentobama'), Bing does a better job segmenting the two words 'president' and 'obama' than the word segment toolkit.
We then count unigrams, bigrams, trigrams, and return: comb. of top 3 unigrams, comb. of 2nd/3rd unigrams top 1 bi/tri-grams if counts reach a threshold Example: #OREvsAZ ? [?orevsaz?, ?oregon?, ?arizona?, ?pac12 championship?]

3.3 Web Search and Selection

3.4 Text Summarization

4. EVALUTATION

At this time, we are not aware of any available "gold standard" dataset for hashtag meanings. As time allows, we will attempt to create such a data set by using a variety of twitter hashtags, with pools defined by mean hashtag popularity, so as to thoroughly test our retrieval pipeline. We expect that scale will be the main weakness of this test dataset. We will also provide a mechanism within the plugin for users to report their search results as inaccurate. This should help with evaluation of the model once it is deployed.
The main contribution of our work is a novel method of extracting meaning from hashtags in order to aid in microblogging post understanding and document recommendation.

5. CONCLUSIONS AND FUTURE WORK