Improving Hashtag Comprehension with Search and Text Extraction

Project Proposal

Jack Lanchantin
University of Virginia
85 Engineer's Way
Charlottesville, VA
22904-4740
lanchantin@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 micro-blogging sites 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 directly 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 the hashtag) on Google, 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 selecting a certain tweet that best explains the hashtag, or by summarizing a large amount of tweets with the same hashtag.

2. RELATED WORK

The research surrounding hashtags covers a variety of topics including retrieval[1], prediction[2][3] 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 corporate sponsorship of individual tags.

Currently, there are a few websites (e.g. tagdef.com, tagboard.com) which attempt to explain hashtags by crowdsourcing definitions. However, these websites do not do a great job defining new hashtags, and they require usersâĂŹ input to define the meaning of hashtags. As theyâĂŹre not so popular among twitter users, the content quality is not as good as the other cloud-sourcing platforms like Wikipedia or Urban Dictionary.

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 EVALUATION

Our implementation plan is as follows: Use the Twitter

API to crawl posts with specific hashtags, implement our three techniques to better understand the hashtag, summarize the hashtag as simply as possible, and display our results as a browse plugin when a user selects a certain hashtag.
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.

3.1 References

4. REFERENCES

[1] Efron, M. Hashtag retrieval in a microblogging environment. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (2010), ACM.

[2] Khabiri, Elham, J. C., and Kamath, K. Y. Predicting semantic annotations on the real-time web. In *Proceedings of the 23rd ACM conference on Hypertext and social media* (2012), ACM.

[3] Weston, James, S. C., and Adams, K. #tagspace: Semantic embeddings from hashtags. In *Empirical Methods in Natural Language Processing (EMNLP) Conference* (2014), EMNLP.