Electronics and Computer Science
Faculty of Physical Sciences and Engineering
University of Southampton

Jamie Davies
25th November 2013

Enhanced Content Analysis and Information Retrieval Using Twitter Hashtags

A project report submitted for the award of
MEng Computer Science

Supervisor:
Dr. Nick Gibbins

Examiner:
Dr. Klaus-Peter Zauner

Abstract

One of the key characteristics of Twitter and other microblogging platforms is the use of 'hashtags' — topical/categorical annotations provided by the authors of the posts (tweets) themselves. This flexible system was designed for the effective organisation and searching of tweets, but with Twitter facing an ever-increasing number of users and tweets it is hard for users to keep track of the vast number of hashtags in popular use. This results in data from the hashtags being fragmented and inaccurate due to the users making poor or uninformed hashtag choices.

If users are presented with a choice of relevant hashtags when writing a tweet, they are more likely to publish tweets with accurate tag data. This project aims to create an intelligent hashtag recommendation tool to raise the information gain from hashtags. However, whilst such a system could improve the quality of the hashtag data for future tweets, tweets that have already been published will remain untouched by the system. Thus, the system will be extended to also retrofit hashtags to published tweets — allowing for tweets to appear in search results for a particular hashtag even if they don't actually contain the hashtag in question.

Contents

1 Project Goals

This project will create a system that aims to support and enrich the information provided by hashtags on Twitter[1]. It will use a combination of different machine-learning techniques to examine and classify the topics and concepts behind the hashtags and in doing so, be able to suggest suitable hashtags for tweets that are relevant to their content. This will allow users to make a better choice of hashtag when writing tweets, and therefore refine the information that they provide.

However, as suggesting better hashtags will only improve the information gain from future tweets, the system will be extended to provide a context-aware tweet search facility. This will enable users to search for a particular hashtag, and instead of only returning tweets containing that hashtag (as current systems do), it will also provide tweets that are contextually relevant to the search term but do not contain that given hashtag.

1.1 Requirements

There are two main requirements the system in this project is aiming to fulfill. It must:

1. Allow users to compose and publish tweets whilst suggesting hashtags relevant to the content of their tweets.
2. Allow users to search for a hashtag and view related tweets, including those that don't contain that hashtag.

These points can be further expanded into a series of functional and non-functional requirements, concluded from the background research.

1.1.1 Functional Requirements

- The system must allow the user to log in and publish tweets to their Twitter account.
- The system must provide hashtag recommendations as the user is creating a tweet.
- The system must perform a hashtag search through a large dataset of tweets and return all relevant tweets, including those that do not contain the search query.

[1] `www.twitter.com`

- The system must use information from a large dataset of tweets to generate a model representing each hashtag.
- The system must be able to compare tweets against its representational hashtag models.
- *Optional:* The system must be able to update its classification models using information from the live Twitter stream.
- *Optional:* The system must provide probabilities for how likely a hashtag is to be related to a tweet.

1.1.2 Non-Functional Requirements

- The system must be accessible via a web interface.
- The system must be responsive and easy to use.
- The system must be able to perform searches quickly.
- The system must be able to make hashtag recommendations quickly.
- The system must be able to produce visualisations to provide an easy way to interpret the hashtag recommendations/assignments.
- *Optional:* The system must be accessible via mobile web browsers.

1.2 Research Questions

This system will provide a new way to gain further insight into how people use hashtags on Twitter, and in doing so, answer the following questions:

- Are certain types of tweet/hashtag easier to classify than others?
- Is it possible to make relevant hashtag suggestions using just the tweet text itself, or is other metadata needed to make the recommendations useful?

2 Background and Literature Review

Some text to give an introduction to the different areas of research.

2.1 Recommendation Systems

Traditional recommendation systems are in place all over the web today. From music discovery services (such as Last.fm[2]) to suggested purchases on retail sites (like that in place at Amazon[3], these systems are all personalised recommendation engines that take an individual user's preferences and use them to provide suggestions tailored to that user.

2.1.1 Collaborative Filtering

Most personalised recommendation systems employ a set of techniques known as collaborative filtering. These techniques were first coined by Goldberg et al. (1992), where a system named *Tapestry* was created that allowed people to attach annotations to documents, and then use that information to filter the documents for other users.

One common implementation of collaborative filtering is the so-called "user-to-user" approach. "User-to-user" collaborative filtering works by taking the preferences of a user A, and finding a small subset of other users in the system that have similar preferences. For each user B in the subset any items that B has adopted that A hasn't are added to a ranked list of suggestions. A is now more likely to adopt items in the list than the items of another random person (Schafer et al., 2001).

2.1.2 Content-Based Recommendation

Another approach to provide relevant recommendations to a user is the use of content-based recommendation systems. This is a type of system that recommends items relevant to other items by comparing the details and descriptions of the items themselves. This can be extended to suggest items for a user by comparing their preferences with the descriptions of the items (Pazzani and Billsus, 2007).

[2] `www.last.fm`

[3] `www.amazon.co.uk`

2.1.3 Relevance Feedback

2.2 Hashtag Recommendation Research

2.3 Research Utilising Hashtag Data

3 Report on Technical Progress

12pt text for main body.

4 Plan of Remaining Work

12pt text for main body.

References

Goldberg, D. et al. (1992). ‘Using Collaborative Filtering to Weave an Information Tapestry’. *Commun. ACM* 35.12, pp. 61–70.

Pazzani, M. and Billsus, D. (2007). ‘Content-Based Recommendation Systems’. *The Adaptive Web*. Ed. by P. Brusilovsky et al. Vol. 4321. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 325–341.

Schafer, J. B. et al. (2001). ‘E-Commerce Recommendation Applications’. *Data Min. Knowl. Discov.* 5.1-2, pp. 115–153.